2018

Haplotype sharing analysis in maize ex-PVP germplasm and relationships with key founders

Stephanie Coffman
Iowa State University

Haplotype sharing analysis in maize ex-PVP germplasm and relationships with key founders

by

Stephanie Coffman

A thesis submitted to the graduate faculty

in partial fulfillment of the requirements for the degree of

MASTER OF SCIENCE

Major: Genetics

Program of Study Committee:
Thomas Lübberstedt, Major Professor
Matthew Hufford
Carson Andorf

The student author, whose presentation of the scholarship herein was approved by the program of study committee, is solely responsible for the content of this thesis. The Graduate College will ensure this thesis is globally accessible and will not permit alterations after a degree is conferred.

Iowa State University

Ames, Iowa

2018

TABLE OF CONTENTS

Page

LIST OF FIGURES

Page

LIST OF TABLES

Page

NOMENCLATURE

Ex-PVP	Expired Plant Variety Protection
IBM	Intermated B73 x Mo17 Population
Indel	Insertions or Deletions
NSS	Non Stiff-Stalk
OPV	Open-Pollinated Variety
QTL	Quantitative Trait Loci
SNP	Single Nucleotide Polymorphism
SS	Stiff-Stalk

ACKNOWLEDGMENTS

I would like to express my thanks to all of those who have been there to support me through this five-year journey. Thank you to my advisor Thomas Lübberstedt for his guidance, support and patience. I would also like to thank Matthew Hufford and Carson Andorf for their guidance and discussions throughout my project. I would like to sincerely thank my colleagues at DowDuPont for believing in me and supporting me. The support from DowDuPont was absolutely crucial to my completion of this program. Thank you specifically to Pedro Hermon, Ed Bruggemann and Laura Mayor for their sincere support when I started this journey and to Andy Baumgarten, Dean Podlich and Justin Gerke for their mentoring and guidance throughout my studies.

In addition, I would like to thank my family, friends and in particular, my husband for sticking with me through the ups and downs that come with working full-time while in a graduate program. I look forward to being able to go on more vacations! Lastly, I would like to thank music and coffee (which I didn't start drinking until grad school) for making the stressful times a little less stressful.

ABSTRACT

The advent of double cross hybrids and eventually single cross hybrids in maize stimulated formation of private breeding programs to develop elite inbred parents. Maize breeding programs in North America have progressed over the last century, driving the formation and evolution of heterotic groups. Modern day maize hybrids are exclusively developed using proprietary inbreds. The Plant Variety Protection Act was passed in 1970 by the U.S. Congress and provides 20 years of legal protection for inbreds and varieties developed for many species. Maize lines with expired Plant Variety Protection (ex-PVP) represent germplasm that is the foundation of many industry seed industry companies. The genetic content of these inbreds can be related back to founder lines through pedigree, molecular data or a combination of the two. Seed companies started with similar genetics from founder lines and have independently used these to evolve their own proprietary germplasm. The structure of this germplasm continues to evolve as breeding programs experiment with different line crosses within each company. Understanding the relationships between ex-PVPs and founder lines will provide insight into the haplotype and heterotic group structure of industry germplasm.

In this study, we utilize high-density SNP data to generate high resolution identity-by-state haplotypes (IBS) haplotypes for 212 maize inbred lines. Among these 212 inbred lines are 157 ex-PVPs registered 1976-1992 and 55 public inbreds relevant for PVP germplasm. These lines include ex-PVPs from the major seed industry companies DowDuPont, Monsanto and Syngenta as well as 12 key founders identified through literature review. We summarize haplotype structure and diversity among these 212 inbreds as well as haplotype sharing between the ex-PVPs and the 12 key founders. We

find that more than 75% of the haplotypes present in these ex-PVPs are shared with at least one of the 12 key founders and the trends in haplotype sharing with founders by company are similar to previous pedigree-based studies. We summarize genome-wide and region-specific haplotype differences among companies and heterotic groups. To facilitate exploration of the haplotype data, a visualization tool was developed using the Shiny framework in R. We summarize this visualization framework using a subset of the 212 inbreds to demonstrate visualization of genome-wide haplotypes. Together, the results from this study demonstrate how haplotype sharing analysis can be utilized to characterize germplasm diversity and provide additional insight into the breeding history of commercial maize.

CHAPTER 1. INTRODUCTION

Hybrid maize history

The evolution of maize (*Zea mays* L.) into one of the most economically important crops worldwide has a rich history. G.H. Shull and E.M. East independently reported on inbreeding depression and its effect on maize hybrid vigor (East, 1908; Shull, 1908). Through crossing of open-pollinated varieties (OPVs) which had been self-pollinated, they showed the F_1 offspring were more vigorous than the initial OPVs. Shull coined the term "heterosis" to describe this phenomenon. Production of seed from single-cross hybrids was initially challenging due to the weak nature of the inbred parents and was not economically feasible for the average farmer (Tracy & Chandler, 2006). Intrigued by their research, D.F. Jones trained with E.M. East and described a method to generate double-cross hybrids through crossing of single-cross hybrids developed from different OPVs (Mangelsdorf, 1975). This approach took advantage of heterosis and the resulting hybrids produced more seed. The prospect of decreased seed production costs and higher yielding hybrids spurred breeder interest in identifying inbred lines, which produced superior hybrids when crossed. As a result, the 1930s and 1940s saw a shift to use of double-cross hybrids over traditional OPVs. Over time, breeders became more proficient at developing strong inbred parents making it feasible for higher yielding single-cross hybrids to replace double-cross hybrids in the 1960s (Troyer, 1999).

Some of the early lineages important in development of the first commercial hybrids were derived from the most successful and widely adapted OPVs and played a role in the development of heterotic patterns (Troyer, 1999). The concept of heterotic patterns began taking hold in the 1960s and 1970s and continued development of successful inbred parents

can be attributed to the formation and divergence of heterotic groups (Tracy & Chandler, 2006). The most widely known heterotic pattern consists of crosses between the heterotic groups stiff-stalk (SS) and non-stiff-stalk (NSS) which were primarily derived from strains of the OPVs Reid Yellow Dent and Lancaster Sure Crop, respectively. Inbred lines tend to be developed from crosses within a heterotic group and are tested in hybrid combination with inbred lines from the opposite heterotic group. Lines that were pivotal in the development of breeding groups, such as these heterotic groups, are termed 'founders' and typically describe the earliest known recorded ancestral genotypes for a given germplasm (Zhou et al., 2000).

Privatization of breeding programs

The seed industry has undergone many changes over the last century including increased privatization of breeding programs and the accelerated use of proprietary inbreds in hybrid development that occurred in the 1980s (Darrah & Zuber, 1986). The transition from use of public to entirely proprietary inbreds (Mikel, 2008) in commercial hybrids underscores the impact of private seed companies in driving the success of North American maize hybrids. In 1970, the Plant Variety Protection Act (PVPA) was passed by U.S. Congress as a means for breeders to protect their innovation. This legal protection expires after 20 years. Once the protection expires, the lines become available to the public. Maize lines with expired Plant Variety Protection (ex-PVP) represent germplasm that is the foundation of many industry seed companies.

Haplotype sharing analysis

A haplotype can be defined as a set of linked alleles, which are inherited together. Haplotypes can provide additional information compared to single-nucleotide polymorphisms (SNPs) because there are more possible combinations of alleles. Haplotypes across individuals are identical-by-state (IBS) when they contain the same allelic

information. Haplotypes have been shown to improve accuracy of genome-wide association studies (GWAS), genomic prediction (Ferdosi et al., 2016; Schrag et al., 2007), mapping of quantitative trait loci (QTL; Kebede et al., 2016; Lu et al., 2010), assessment of population structure and depiction of relationships between individual samples (Fang et al., 2014; Gattepaille & Jakobsson, 2012; Haasl & Payseur, 2011; Lawson et al., 2012; Ralph & Coop, 2013).

Haplotypes are typically generated from the ordered combinations of SNP alleles using a linkage disequilibrium (LD) or window-based approach. In a LD-based approach, an algorithm identifies the natural recombination breakpoints among haplotypes in a population. This leads to haplotypes of varying physical lengths. A window-based approach divides the genome into equal sized windows and haplotype groups are identified within each window. Simple clustering and similarity algorithms (Gusev et al., 2009; Purcell et al., 2007; Swarts et al., 2014; Ward, 1963) can be applied to identify haplotypes. Sets of ordered SNP alleles that cluster together based on some similarity threshold in a genomic region would be considered IBS. More complex probability-based approaches which incorporate recombination data can also be used (Browning & Browning, 2009; Daly et al., 2001).

Haplotype sharing has been used as a measure of relatedness and can reveal regions of selection and variation in diversity present within germplasm (Fang et al., 2014; Hufford et al., 2013; Poets et al., 2016). In maize, a few studies have identified specific segments of IBS haplotype sharing between a small number of founders and select maize inbreds (Dell'Acqua et al., 2015; Jiao et al., 2012; Romero-Severson et al., 2001; Wu et al., 2016). Haplotype sharing analyses enable identification of shared regions across individuals and can provide context to breeding history and diversity.

Visualization of haplotype structure

A visualization of population structure is an important tool for researchers. Population structure is frequently assessed through principal components analysis (PCA) and generation of STRUCTURE and fastStructure plots (Pritchard et al., 2000; Raj et al., 2014). These types of analyses produce clusters and groupings of individuals allowing the researcher to identify individuals that have common underlying genomic features. At the nucleotide level, color coding of SNP data with tools such as Flapjack (Milne et al., 2010) can provide insight into the haplotype structure of a group of individuals. Tools such as Haploview (Barrett et al., 2005) and GEVALT (Davidovich et al., 2007) enable linkage disequilibrium (LD) based haplotype construction and identification of tag SNPs useful in downstream analyses but lack visualization to provide context to the haplotypes. Tools that do provide haplotype visualization are often targeted for region-specific visualization rather than whole-genome visualization. Some examples include inPHAP (Jäger et al., 2014), Haplostrips (Marnetto et al., 2017) and HaploForge (Tekman et al., 2017), which is modeled after HaploPainter (Thiele & Nürnberg, 2005) and generates pedigree-based haplotype visualizations for user-defined regions.

Research objectives

The research objectives of this study were to i) evaluate haplotype structure in germplasm that has driven the success of the seed corn industry and ii) summarize haplotype sharing across companies and heterotic groups to gain insight into the breeding history and diversity of commercial maize as it relates to key founders. High-density SNP data from a publicly available dataset was utilized to generate high-resolution haplotypes through a window-based haplotype approach. Overall population structure and haplotype diversity were assessed. Haplotype sharing between ex-PVPs and key maize founders was examined

across seed industry companies and heterotic groups. The results from this study demonstrate application of haplotype sharing analysis in a study of breeding history and germplasm diversity. A tool was developed using the R Shiny framework to visualize haplotypes as a result of the analysis process for the main study objectives. The objective of developing this interface was to enable visualization of genome-wide haplotypes and comparisons of haplotype structure across individuals.

Thesis formatting

This thesis contains two manuscripts in preparation for journal submission. Chapter Two is in the standard format for publication in *Theoretical and Applied Genetics* and summarizes haplotype structure and diversity among ex-PVP lines in relation to key founders. My contributions to Chapter Two include project design, sample and locus selection and analyses, diversity and haplotype analyses, interpretation and summarization of results and preparation of the manuscript. Chapter Three is in the standard format for publication in *BMC Bioinformatics*. This chapter introduces a tool developed using the R Shiny framework for visualization of haplotype data and comparison of haplotypes across individuals. This tool was developed to provide an interactive way for researchers to explore haplotype datasets and examine similarities and differences in genomic regions of interest. My contributions to Chapter Three include building the sample haplotype dataset, writing the scripts to develop the Shiny graphical user interface and visualizations and preparation of the manuscript.

References

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics, 21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Browning, B. L., & Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*, *84*(2), 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, *29*(2), 229–232. https://doi.org/10.1038/ng1001-229

Darrah, L. L., & Zuber, M. S. (1986). 1985 United States Farm Maize Germplasm Base and Commercial Breeding Strategies 1. *Crop Science*, *26*(6), 1109–1113. https://doi.org/10.2135/cropsci1986.0011183X002600060004x

Davidovich, O., Kimmel, G., & Shamir, R. (2007). GEVALT: An integrated software tool for genotype analysis. *BMC Bioinformatics*, *8*, 36. https://doi.org/10.1186/1471-2105-8-36

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., … Pè, M. E. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. *Genome Biology*, *16*, 167. https://doi.org/10.1186/s13059-015-0716-z

East, E. M. (1908). Inbreeding in corn. *Connecticut Agric. Exp. Stn. Rep., 1907*, 419–428.

Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J., & Morrell, P. L. (2014). Two genomic regions contribute disproportionately to geographic differentiation in wild barley. *G3: Genes, Genomes, Genetics*, *4*(7), 1193–1203. https://doi.org/10.1534/g3.114.010561

Ferdosi, M. H., Henshall, J., & Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. *Genetics Selection Evolution*, *48*, 75. https://doi.org/10.1186/s12711-016-0253-6

Gattepaille, L. M., & Jakobsson, M. (2012). Combining Markers into Haplotypes Can Improve Population Structure Inference. *Genetics*, *190*(1), 159–174. https://doi.org/10.1534/genetics.111.131136

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., … Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*(2), 318–326. https://doi.org/10.1101/gr.081398.108

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. https://doi.org/10.1038/hdy.2010.21

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The Genomic Signature of Crop-Wild Introgression in Maize. *PLOS Genetics*, *9*(5), e1003477. https://doi.org/10.1371/journal.pgen.1003477

Jäger, G., Peltzer, A., & Nieselt, K. (2014). inPHAP: Interactive visualization of genotype and phased haplotype data. *BMC Bioinformatics*, *15*, 200. https://doi.org/10.1186/1471-2105-15-200

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., … Lai, J. (2012). Genome-wide genetic changes during modern breeding of maize. *Nature Genetics*, *44*(7), 812–815. https://doi.org/10.1038/ng.2312

Kebede, A. Z., Woldemariam, T., Reid, L. M., & Harris, L. J. (2016). Quantitative trait loci mapping for Gibberella ear rot resistance and associated agronomic traits using genotyping-by-sequencing in maize. *Theoretical and Applied Genetics*, *129*(1), 17–29. https://doi.org/10.1007/s00122-015-2600-3

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., … Xu, Y. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences*, *107*(45), 19585–19590. https://doi.org/10.1073/pnas.1006105107

Mangelsdorf, P. C. (1975). Donald Forsha Jones (1890-1963). In *Biographical Memoirs* (Vol. 46, pp. 135–156). National Academy of Science. Retrieved from https://doi.org/10.17226/569

Marnetto, D., Huerta-Sánchez, E., & Price, S. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods in Ecology and Evolution*, *8*(10), 1389–1392. https://doi.org/10.1111/2041-210X.12747

Mikel, M. A. (2008). Genetic Diversity and Improvement of Contemporary Proprietary North American Dent Corn. *Crop Science*, *48*(5), 1686. https://doi.org/10.2135/cropsci2008.01.0039

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., … Marshall, D. (2010). Flapjack—graphical genotype visualization. *Bioinformatics*, *26*(24), 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Poets, A. M., Mohammadi, M., Seth, K., Wang, H., Kono, T. J. Y., Fang, Z., … Morrell, P. L. (2016). The Effects of Both Recent and Long-Term Selection and Genetic Drift Are Readily Evident in North American Barley Breeding Populations. *G3: Genes, Genomes, Genetics*, *6*(3), 609–622. https://doi.org/10.1534/g3.115.024349

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based

linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ralph, P., & Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. *PLOS Biology*, *11*(5), e1001555. https://doi.org/10.1371/journal.pbio.1001555

Romero-Severson, J., Smith, J. S. C., Ziegle, J., Hauser, J., Joe, L., & Hookstra, G. (2001). Pedigree analysis and haplotype sharing within diverse groups of Zea mays L. inbreds. *Theoretical and Applied Genetics*, *103*(4), 567–574. https://doi.org/10.1007/PL00002911

Schrag, T. A., Maurer, H. P., Melchinger, A. E., Piepho, H.-P., Peleman, J., & Frisch, M. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theoretical and Applied Genetics*, *114*(8), 1345–1355. https://doi.org/10.1007/s00122-007-0521-5

Shull, G. H. (1908). The Composition of a Field of Maize. *Journal of Heredity*, *os-4*(1), 296–301. https://doi.org/10.1093/jhered/os-4.1.296

Swarts, K., Li, H., Navarro, R., Alberto, J., An, D., Romay, M. C., … Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome*, *7*(3). https://doi.org/10.3835/plantgenome2014.05.0023

Tekman, M., Medlar, A., Mozere, M., Kleta, R., & Stanescu, H. (2017). HaploForge: a comprehensive pedigree drawing and haplotype visualization web application. *Bioinformatics*, *33*(24), 3871–3877. https://doi.org/10.1093/bioinformatics/btx510

Thiele, H., & Nürnberg, P. (2005). HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, *21*(8), 1730–1732. https://doi.org/10.1093/bioinformatics/bth488

Tracy, W. F., & Chandler, M. A. (2006). The Historical and Biological Basis of the Concept of Heterotic Patterns in Corn Belt Dent Maize. In K. R. Lamkey & M. Lee (Eds.), *Plant Breeding: The Arnel R. Hallauer International Symposium* (pp. 219–233). Blackwell Publishing. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470752708.ch16/summary

Troyer, A. F. (1999). Background of U.S. Hybrid Corn. *Crop Science*, *39*(3), 601–626. https://doi.org/10.2135/cropsci1999.0011183X003900020001x

Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wu, X., Li, Y., Fu, J., Li, X., Li, C., Zhang, D., … Wang, T. (2016). Exploring Identity-By-Descent Segments and Putative Functions Using Different Foundation Parents in Maize. *PLOS ONE*, *11*(12), e0168374. https://doi.org/10.1371/journal.pone.0168374

Zhou, X., Carter, T. E., Cui, Z., Miyazaki, S., & Burton, J. W. (2000). Genetic Base of Japanese Soybean Cultivars Released during 1950 to 1988. *Crop Science*, *40*(6), 1794–1802. https://doi.org/10.2135/cropsci2000.4061794x

CHAPTER 2. HAPLOTYPE STRUCTURE DIVERSITY AMONG COMMERCIAL MAIZE EX-PVP LINES IN RELATION TO KEY FOUNDERS

Article in preparation for publication in *Theoretical and Applied Genetics*

Stephanie M. Coffman[1,2], Matthew B. Hufford[3], Carson M. Andorf[4], Thomas Lübberstedt[5]

Abstract

Key message

High-density haplotype analysis revealed significant haplotype sharing between ex-PVPs registered from 1976-1992 and key maize founders and uncovered similarities and differences in patterns of sharing by company and heterotic group.

Abstract

Proprietary inbreds developed by private seed industry companies have been the major source for driving genetic gain in successful North American maize hybrids for decades. Much of the history of industry germplasm can be traced back to key founder lines, some of which were pivotal in the development of prominent heterotic groups. Previous studies have summarized pedigree-based relationships, genetic diversity and population structure among ex-PVPs, however, little is known about the extent of haplotype sharing between historical founders and ex-PVP inbreds. A better understanding of the relationships between founder lines and ex-PVPs would provide insight into the haplotype and heterotic group structure among industry germplasm. We performed high-density haplotype analysis with 11.3 million SNPs on a selection of 212 maize inbreds, which included 157 ex-PVP inbreds registered 1976-1992 and 55 public lines relevant to PVPs. Among these lines were

[1] Primary Author
[2] Systems and Innovation for Breeding and Seed Products, DowDuPont, Johnston, Iowa
[3] Department of Ecology, Evolution & Organismal Biology, Iowa State University, Ames, Iowa
[4] USDA-ARS Corn Insects and Crop Genetics Research Unity, Iowa State University, Ames, Iowa
[5] Department of Agronomy, Iowa State University, Ames, Iowa

12 key founders identified in literature review: 207, A632, B14, B37, B73, LH123HT, LH82, Mo17, Oh43, OH7, PHG39 and Wf9. Our results revealed that 75.8% of the ex-PVP haplotype space is shared with at least 1 of these 12 founder lines and more than half is collectively shared with B73, Mo17 and 207. Quantifiable similarities and contrasts among heterotic groups and major U.S. seed industry companies were also observed. The results from this study provide high-resolution haplotype data on ex-PVP germplasm, confirm founder relationship trends observed in previous studies, uncover region-specific haplotype structure differences and demonstrate how haplotype sharing analysis can be used as a tool to explore germplasm diversity.

Keywords

haplotype, haplotype sharing, ex-PVP, germplasm diversity

Introduction

The North American maize industry has a rich history, which contributes to its success in production of superior hybrids and feeding of a global population. Inbreeding depression and its effect on maize hybrid vigor was independently reported by G.H. Shull and E.M. East and is termed heterosis (East, 1908; Shull, 1908). The 1930s and 1940s saw a shift from traditional open-pollinated varieties (OPVs) to double-cross hybrids as this was the most economically viable seed production approach at the time (Mangelsdorf, 1975). Through selection, breeders developed stronger inbred parents making it feasible for higher yielding single-cross hybrids to replace double-cross hybrids in the 1960s (Troyer, 1999). Continued development of successful inbred parents can be attributed to the formation and divergence of heterotic groups, a concept that began taking hold in the 1960s and 1970s (Tracy & Chandler, 2006). The most widely known heterotic pattern consists of the heterotic groups stiff-stalk (SS) and non-stiff-stalk (NSS). Inbred lines tend to be developed from

crosses within a heterotic group and are tested in hybrid combination with inbred lines from a complementary heterotic group. Lines that were pivotal in the development of these breeding groups are called ‘founders’, a term that typically describes the earliest known recorded ancestral genotypes for a given germplasm (Zhou et al., 2000). The history of heterotic groups in maize can be traced back through pedigree information to founders which were key in their development (Bernardo et al., 2000; J. S. C. Smith et al., 1999).

The transition from use of public to entirely proprietary inbreds (Mikel, 2008) in commercial hybrids underscores the impact of private seed companies in driving the success of North American maize hybrids. In 1970, the Plant Variety Protection Act (PVPA) was passed by U.S. Congress as a means for breeders to protect their innovations. Approved Plant Variety Protection (PVP) of individual maize lines provides legal protection, which expires after 20 years. Once expired, the lines become available to the public. Maize lines with expired Plant Variety Protection (ex-PVP) represent germplasm that is the core of many seed companies. As of June 2017, 386 field corn inbred lines had expired PVP. Company representation among these lines is reflective of the major stakeholders in the industry such as DowDuPont, Monsanto and Syngenta. Seed companies started with similar genetics from founder lines (Troyer, 1999). As seed companies have independently used these founders to evolve their own proprietary germplasm, it should not be expected that founder lines contributed equally to breeding programs across the agricultural industry.

Understanding how founder lines have contributed to the development of ex-PVP lines can provide insight into industry germplasm structure and relationships between pre-commercial and commercial maize lines. Founder lines with significant contributions to PVP lines have been identified through pedigree-based analysis (Mikel, 2008, 2011; Mikel &

Dudley, 2006; S. Smith, 2007). SNP-based analyses have confirmed the importance of these founders in ex-PVP germplasm and identified common heterotic groups (Beckett et al., 2017; van Heerwaarden et al., 2012; Nelson et al., 2008; Romay et al., 2013). The three main heterotic groups in ex-PVP germplasm are stiff-stalk (SS), non-stiff stalk (NSS) and Iodent. Genetic analyses have shown that these heterotic groups are the product of modern breeding rather than a result of historical divergence among landrace founders (van Heerwaarden et al., 2012). Narrowing of ancestral composition in heterotic groups over time has resulted in significantly increased shared haplotype lengths in ex-PVPs compared to their predecessors (van Heerwaarden et al., 2012; Romay et al., 2013). A few studies have identified specific segments of identity-by-state (IBS) haplotype sharing between a small number of founders and select maize inbreds (Dell'Acqua et al., 2015; Jiao et al., 2012; Romero-Severson et al., 2001; Wu et al., 2016).

Despite these research efforts, no comprehensive study assessing and relating identity-by-state (IBS) haplotype sharing between founders and ex-PVP lines to the individual seed industry companies and heterotic groups has been conducted. Large scale genotyping studies on diverse maize lines increased access to higher density SNP data on ex-PVPs and historically important public lines (Bukowski et al., 2018; Romay et al., 2013). Here, we build on previous studies and utilize high-density SNP data to provide new analyses focused on the haplotype structure among ex-PVP lines. Our objectives were to i) evaluate haplotype structure in germplasm that has driven the success of the seed corn industry, and ii) summarize haplotype sharing across companies and heterotic groups to gain insight into the breeding history and diversity of commercial maize as it relates to key founders.

Materials and Methods

Identification of ex-PVPs and public lines

The U.S. PVP Application Status report was downloaded from https://www.ams.usda.gov/services/plant-variety-protection/application-status on June 30, 2017. The report contained 386 field corn lines with expired PVP. Pedigrees were extracted from the PVP applications for a subset of 174 ex-PVPs represented in the MaizeHapMap3.2.1 (Bukowski et al., 2018).

In addition to the ex-PVP lines, 60 inbreds which were identified as ancestors to maize PVPs in previous studies (Beckett et al., 2017; van Heerwaarden et al., 2012; Mikel, 2011) were included. Twelve lines, among these 174 ex-PVPs and 60 public lines, were selected as key founders for this study: 207, A632, B14, B37, B73, LH123HT, LH82, Mo17, Oh43, OH7, PHG39 and Wf9. These lines were selected based on previous studies of PVPs and ex-PVPs which identified important progenitor lines based on pedigree relatedness and population structure analyses (Beckett et al., 2017; van Heerwaarden et al., 2012; Mikel, 2008, 2011; Mikel & Dudley, 2006; Paul T. Nelson et al., 2008; Romay et al., 2013).

Genotypic data

Unimputed SNP calls from the MaizeHapMap3.2.1 (Bukowski et al., 2018) uplifted to B73 AGPv4 were downloaded through CyVerse Data Store. This full dataset contained 1,218 samples and 81,687,392 loci. The estimated error rate for this published dataset is 1-3%. A subset of 260 samples were extracted using BCFtools 1.5 (Li, 2011), representing 234 unique inbreds consisting of the 174 ex-PVPs and 60 public lines.

Filtering of samples and loci

The FilterSiteBuilderPlugin in Tassel 5.0 (Bradbury et al., 2007) was used to remove monomorphic loci, loci with >2 alleles or >95% missing data points and loci with >5%

heterozygous calls. Inbreds with multiple replicates were reviewed to determine if the replicates could be merged. A set of 500,000 loci (50,000 per chromosome) was used for sample replicate analysis. Pairwise comparisons of sample replicates within each inbred were made at overlapping, non-missing loci.

Comparisons of the genotype calls for replicate pairs which showed <95% match rate were examined in greater detail. These discordant replicates were compared to any available parent and/or progeny replicates to identify the most representative sample. If a representative sample could not be identified, then all replicates were discarded.

Sample replicates with >95% match rate were merged as follows, at a given locus:

- If all calls across reps are NA (missing), return NA
- If there is only one unique genotype call across reps, return this genotype
- If there is more than one unique genotype call (e.g. AA and GG) across reps and they occur at equal frequency, return NA
- If there is more than one unique genotype call and one of them occurs at a higher frequency than others, return the most frequent genotype call

Samples that did not pass sample replicate analysis and samples with >80% missing data were removed.

After sample replicate analysis, loci containing indels and loci with >50% missing data points were excluded. A test set of haplotypes was built using Tassel's FILLINFindHaplotypesPlugin. A window size of 2,000 SNPs was selected to improve genome coverage of the output haplotypes and enable haplotype block sizes smaller than the average haplotype length of 5.1 Mb reported in Romay et al. (2013). Minimum number of samples required to form a haplotype group was set to 1 to allow haplotypes unique to a

single line in the results. Maximum diversity was set to 3% to account for sequencing error while maintaining stringency on grouping highly similar lines into the haplotype groups. To account for residual heterozygosity, up to 5% heterozygous calls were allowed in the consensus haplotypes. All other settings used the defaults, which included a maximum of 40% missing allowed in a sample within a haplotype block. Inbreds which did not have at least 50% of the genome assigned to haplotypes were excluded from further analysis.

Estimating genetic positions

Genetic positions were estimated for 39,261 SNPs based on the SNP50 chip used by Gerke et al. (2015) using a map derived from the B73 x Mo17 (IBM) mapping population (Ganal et al., 2011). Physical coordinates for the SNP50 loci on the B73 AGPv3 reference were acquired through the MaizeSNP50 Manifest File available at https://support.illumina.com/array/array_kits/maizesnp50_dna_analysis_kit/downloads.html. The AGPv3 coordinates were converted to B73 AGPv4 coordinates using the Gramene Assembly Converter http://ensembl.gramene.org/Zea_mays/Tools/AssemblyConverter?db=core. Loci with different chromosome assignments or no chromosome assignment on AGPv4 when compared to Gerke et al. (2015) were removed. A custom R script was used to identify 35,951 loci that showed collinearity between physical and genetic positions (Supplementary File S1). Using the 35,951 collinear loci as framework markers, genetic positions were estimated for non-framework loci by linear interpolation as implemented in the 'na.approx' function from the 'zoo' library in R (R Core Team, 2016; Zeileis & Grothendieck, 2005) with rule = 2. Estimated physical ranges for centromeres on B73 AGPv4 were provided by MaizeGDB based on data originally mapped to RefGen v2 (PMID: 19956743). Genetic positions were estimated for these centromeres for display in the plots.

Population structure

Principal components (PC) were estimated using Tassel's PrincipalComponentsPlugin which takes the genotypic data as input. Loci with >25% missing were excluded from the input dataset. Cluster assignments were determined using partitioning around medoids (PAM; Kaufman & Rousseeuw, 2008) and fastStructure (Raj et al., 2014). PAM was run using the 'pam' function from the R library 'cluster' (Maechler et al., 2018). The distance matrix (1-IBS) required for this function was generated using Tassel's DistanceMatrixPlugin. PAM clusters were determined for k values 2 to 10. To determine the optimal k value, the average silhouette width was plotted for each k. A k value of 4 showed a high average silhouette width and fit with the known population structure among ex-PVPs. fastStructure was then ran using default settings and k=4. The four clusters were labeled based on the primary founder in the group – B73, Mo17, 207 or Oh43. Inbreds that had the same cluster assignment by both the PAM and fastStructure methods were given a heterotic group assignment. Samples in the B73 cluster were assigned to the stiff-stalk (SS) heterotic group while samples in the Mo17, 207 and Oh43 clusters were assigned to the non-stiff stalk (NSS), Iodent and Oh43-type heterotic groups, respectively.

Haplotype analysis

Samples were assigned to haplotype groups within fixed SNP windows using Tassel's FILLINFindHaplotypesPlugin (Swarts et al., 2014) with the same parameters, which were used when building the test set of haplotypes in the sample filtering process. These parameters were a SNP window size 2,000, minimum number of samples to form a haplotype group 1, maximum diversity 3%, maximum heterozygosity 5% and maximum missing in a sample within a haplotype block 40%. Each fixed SNP window will be referred to as a haplotype block.

Two metrics were used to review haplotype diversity. The first metric, haplotype diversity (*H*), was computed using the method described by Nei (1987) and is shown by the equation below where x_i is the relative frequency of the haplotype within the haplotype block and N is the number of samples.

$$H = \frac{N}{N-1}\left(1 - \sum_i x_i^2\right)$$

Second, Manhattan distance between consensus haplotypes within haplotype blocks was used as a second metric and computed using the 'dist' function in R. The consensus haplotype for each haplotype group within each haplotype block was obtained from the FILLINFindHaplotypes plugin output. Heterozygous calls present in the consensus haplotypes were converted to missing and remaining homozygous calls were converted to 0,1 format, representing the two homozygous classes.

Identity-by-state (IBS) haplotype sharing was computed for ex-PVP and founder inbred pairs as the proportion of haplotype space (physical or genetic) in which the ex-PVP belonged to the same haplotype group as a given founder. First, for each ex-PVP, the total Mb and cM that received a haplotype assignment by the FILLINFindHaplotypes plugin were determined. This was termed the haplotype space. Second, haplotype blocks, where the ex-PVP and given founder line received the same haplotype group assignment were identified. Third, the amount of physical and genetic spaces shared by the haplotype blocks was computed and divided by the total haplotype space (physical or genetic) for that ex-PVP. To determine % IBS haplotype sharing between a group of samples and a given founder, the individual IBS haplotype sharing proportions were simply averaged.

Results

Filtered samples and loci

Thirty-five pairs of sample replicates representing 20 inbreds were compared. SNP calls compared between replicate pairs had a match rate greater than 97% in all but six cases (Supplementary File S2). Widespread mismatches were observed between replicates for the inbreds A619, C103, H99, W153R and W117, while the ND246 replicates showed mismatches localized to specific chromosomal regions. Use of pedigree information in combination with parent-progeny comparisons of SNP data resolved discordance in the case of C103 (Figure A 2.1), A619 and H99. A single replicate was selected for each of these three inbreds. Comparisons of the replicates to available parent and/or progeny replicates were inconclusive for ND246, W153R and W117. Consequently, all replicates for those inbreds were removed from the dataset. Genotypic data for replicate pairs with >97% match were merged. After collapsing and selection of replicates, two additional samples with >80% missing data were removed. Samples that initially had one replicate in the dataset are assumed to be representative of the breeding source. The remaining dataset contained 229 samples and 31,309,187 loci.

In addition to the loci and samples removed in the sample replicate analysis process, 20,012,498 loci which had >50% missing data or contained indels across the 229 samples were excluded. A test set of haplotypes built with this dataset revealed 17 inbreds with <50% genome coverage in genetic space. Removal of these 17 inbreds reduced the genotypic dataset to 212 samples and 11,296,689 loci. The resulting loci were well distributed across the physical genome with very few regions showing decreased coverage. Regions with decreased coverage across the physical genome most often occurred around the centromeres.

Median missing data was 38.1% and 38.7% for samples and loci, respectively. Missing data was <5% for key founders 207, A632, B73, Mo17 and Oh43 due to higher sequencing depth in the source data and collapsing of multiple sample replicates. Among the 212 samples were 55 public lines, including the 12 founders, and 157 ex-PVPs. The 55 public lines had a median of 20.8% missing data and the ex-PVPs had a median of 42.85% missing data. Although imputation could have been used with this dataset, the reduced amount of missing data for the public lines was sufficient to anchor most of the genome space in the ex-PVPs to haplotype groups that could be related to these public lines.

Company representation among the 157 ex-PVPs is shown in Figure 2.1. Mergers and acquisitions have resulted in the integration of germplasm from multiple legacy programs into three main companies: DowDuPont, Monsanto and Syngenta. Haplotype information will be primarily summarized based on these three companies. As such, the dataset contains 65, 63 and 15 ex-PVPs for the three companies, respectively. The full list of samples with additional metadata can be found in Supplementary File S3.

Population structure

Removal of loci with >25% missing data for the population structure analyses resulted in a dataset with the 212 samples and 1,143,283 loci. PC analysis revealed three main clusters relating to the known SS, NSS and Iodent heterotic groups and a 4th cluster consisting of Oh43-type lines (Figure 2.2). Proportions of the total variance explained by PC1, PC2 and PC3 were 7.4%, 5.6%, and 3.8%, respectively. PC1 separated out the SS and NSS clusters, PC2 separated out the Iodent cluster and PC3 separated out the Oh43 cluster.

155 of 212 inbreds had the same cluster assignment by fastStructure and PAM. These 155 inbreds consisted of 66 SS, 26 NSS, 31 Iodent and 32 Oh43-type lines. Heterotic group assignments for inbreds labeled as SS, NSS and Iodent agreed with previous publications

(Beckett et al., 2017; van Heerwaarden et al., 2012; Mikel, 2011). These studies did not emphasize an Oh43-like subpopulation. However, a comparison to Nelson et al. (2008) confirmed cluster assignment for several Oh43-type lines.

Of the 155 inbreds assigned to a heterotic group, 126 were ex-PVPs. Among these were 46 DowDuPont, 58 Monsanto and 12 Syngenta ex-PVPs. Heterotic group assignments among the ex-PVPs were 52 SS, 21 NSS, 31 Iodent and 22 Oh43-types. None of the DowDuPont ex-PVPs were given a NSS heterotic group assignment.

Haplotype structure and diversity

The haplotype blocks, as defined by the fixed SNP windows, had a median physical size of 337.9 kb and maximum of 5 Mb. In genetic space, the median genetic size was 0.04 cM and the maximum was 5.56 cM. On average there were 23.8 haplotype groups per haplotype block. The average number of haplotype groups in 1 cM bins ranged from 12.2 to 56.0 and from 6.0 to 52.0 for 1 Mb bins (Figure 2.3). The average number of haplotype groups was lower in pericentromeric regions and higher near the ends of the chromosome for all ten chromosomes. A few chromosomes showed noticeable decreases in average number of haplotype groups in the chromosome arms. These regions include, but are not limited to, chromosome 3 at 200 Mb, chromosome 4 at 50 Mb and 210 Mb, chromosome 5 near 40 Mb and chromosome 10 at 100 Mb. Haplotype groups containing one inbred were allowed in haplotype blocks provided the other parameters of the FILLINFindHaplotypes plugin were met. This enabled observation of haplotypes such as those unique to an ex-PVP that were not shared with other ex-PVPs or public lines in this dataset. Excluding haplotype groups that only contained one individual, there was an average of 15.3 haplotype groups per haplotype block. Inbreds had an average of 92.9% of the physical genome and 91.5% of the genome in

genetic space assigned to haplotype groups. Complete haplotype assignments for the 212 inbreds can be found in Supplementary File S4.

The number of haplotype groups present in each heterotic group was reviewed using the heterotic group assignments from the population structure analysis. The total number of haplotype groups containing at least 1 individual from the full set of 212 inbreds was 129,721. At least 1 SS line was present in 45,199 haplotype groups. For NSS, Iodent and Oh43-type this was 36,207, 34,087 and 50,531 haplotype groups, respectively. A number of haplotype groups were also shared between pairs of heterotic groups (Table 2.1). 21.5% of the haplotype groups present among SS and NSS were shared. 31.8% of haplotype groups among SS and Iodent lines were shared and 24.7% were shared between Iodent and NSS. Oh43-type lines shared 27.8%, 31.5% and 31.4% haplotype groups with SS, NSS and Iodent, respectively. All four heterotic groups were represented in 9.3% of the haplotype groups present among these heterotic groups. Sharing of haplotype groups was also present among companies with 43.2% shared between DowDuPont and Monsanto. Syngenta shared 35.0% and 33.9% with DowDuPont and Monsanto, respectively. The percentage of shared haplotype groups with Syngenta may be biased due to low counts of Syngenta ex-PVPs (15 lines) compared to DowDuPont and Monsanto.

Average genome-wide haplotype diversity (*H*) was 0.89 when weighted by block size in cM and 0.82 when weighted by block size in kb. The presence of singleton haplotypes (haplotype groups containing one inbred) did not significantly influence the haplotype diversity metrics computed. *H* values varied across the genome with values ranging from 0.54 to 0.95 by cM (Figure 2.4). Pericentromeric regions tend to show a decrease in *H*, which is more easily observed on the genetic scale. A decrease in *H* among the 212 inbreds on

chromosome 4 from 42-67 Mb corresponds with a region identified by Romay et al. (2013), where longer average haplotype lengths were observed among ex-PVPs. Romay et al. (2013) note that this region is known to contain genes related to selection during domestication and improvement processes (Hufford et al., 2012; Lai et al., 2010). Selective sweeps in ex-PVPs in this region were identified by Jiao et al. (2012).

Consensus haplotypes were constructed for each haplotype group, and the distances between these consensus haplotypes were quantified by Manhattan distance. The average pairwise distance was 0.24 when weighted by block size in cM and 0.25 when weighted by block size in kb. Manhattan distance values varied across the genome ranging from 0.11 to 0.33 per cM (Figure 2.4). Some pericentromeric regions showed lower average Manhattan distance across multiple haplotype blocks indicating regions where the haplotype groups are more similar to one another.

Haplotype sharing between ex-PVPs and 12 key founders

The top founders for each company based on IBS haplotype sharing in genetic space (Table 2.2; Supplementary File S5) were similar to previous studies (Mikel, 2008, 2011; Mikel & Dudley, 2006; S. Smith, 2007). B73 shares more haplotype space with Monsanto and Syngenta ex-PVPs, as compared to DowDuPont, and was the top founder by IBS haplotype sharing among all 157 ex-PVPs studied. B73 was IBS with 25.6% of the genetic haplotype space in Monsanto ex-PVPs and 33.5% with Syngenta ex-PVPs while it only shared 13.6% with DowDuPont ex-PVPs. The DowDuPont ex-PVPs shared an average of 25.4% of the genetic haplotype space with 207, making it the top founder for these DowDuPont ex-PVPs. Mo17 had much lower IBS haplotype sharing with DowDuPont ex-PVPs at 7.4% compared to 19.8% and 23.1% with Monsanto and Syngenta, respectively.

A cumulative IBS haplotype sharing plot (Figure 2.5) showed that the majority of the haplotype space across ex-PVPs for the three main seed industry companies in the U.S. can be accounted by just a few founders. Incremental addition of each founder increases the haplotype space, which can be related to the ex-PVPs, but only to a certain degree. Overall, 75.8% of the combined haplotype space among the 157 ex-PVP lines is related to at least one of the 12 founder lines used in this study. Both Monsanto and Syngenta have a higher proportion of the haplotype space IBS with those 12 founders at 80.9% and 76.7%, respectively, while DowDuPont shares 69.6%. The key representatives of the SS, NSS and Iodent heterotic groups – B73, Mo17 and 207, respectively – accounted for 51.1% of the haplotype space across the ex-PVP lines. By company this was 43% for DowDuPont, 55% for Monsanto and 62.7% for Syngenta.

It is important to note that the 12 founders selected in this study are not completely distinct from one another. Figure 2.6 shows a comparison of haplotype group assignments among the 12 founders. 26.5% of the genetic haplotype space in B73 is shared with B14 and 20.7% with B37. 26.4% of the genetic haplotype space in LH82 is shared with 207. 47.4% of the genetic haplotype space in PHG39 is shared with B37. 207 shares 5.1%, 6.5%, 5.0% and 7.6% of its genetic haplotype space with B14, B73, Mo17 and Wf9, respectively. B37, a cycle 0 BSSS derivative shares 12.1% of its genetic haplotype space with B73, a cycle 5 BSSS derivative.

Region-specific differences in IBS haplotype sharing

Haplotype structure is readily observed among DowDuPont and Monsanto ex-PVPs by plotting haplotype group assignments across the inbreds (Figure 2.7). Haplotypes representative of the key founders for each of the four heterotic groups are prevalent among the ex-PVPs in those heterotic groups. Among the SS lines in Figure 2.7, B73 and B14

haplotypes are represented by dark purple and yellow colored blocks, respectively. Haplotype blocks shared with Mo17, 207 and Oh43 are represented by green, blue and red blocks, respectively. Composite views of the IBS haplotype assignments by company highlight the variation in haplotype structure within these ex-PVPs (Figure A 2.3; Figure A 2.4). Regions of similarity and difference within and across heterotic groups and between companies are evident. In particular, Mo17 shows greater haplotype sharing on average across the entire genomes rather than just specific regions among Monsanto ex-PVPs compared to DowDuPont ex-PVPs. Similarly, 207 shows greater haplotype sharing across the entire genome with DowDuPont ex-PVPs compared to Monsanto ex-PVPs. The decrease in *H* observed in the chromosome 4 domestication region is reflected by the strong presence of B73 and Mo17 haplotypes in this region for both DowDuPont and Monsanto ex-PVPs.

IBS haplotype sharing with a given founder is not constant across the genome. Average IBS haplotype sharing with a given founder was quantified across physical and genetic windows for specific groups of inbreds. For example, B73 has higher haplotype sharing with Monsanto and Syngenta ex-PVPs not just at the whole genome level but at nearly all physical and genetic windows as compared to DowDuPont (Figure 2.8). This may be somewhat biased due to the counts of DowDuPont inbreds assigned to heterotic groups other than SS compared to Monsanto and Syngenta. 44.4% of Monsanto ex-PVPs and 46.4% of Syngenta ex-PVPs were assigned to the SS heterotic group compared to 23.1% for DowDuPont ex-PVPs. Nonetheless, it highlights the strong presence of B73 haplotypes in these Monsanto and Syngenta ex-PVPs. The similarities and differences in haplotype sharing with B73 are more evident when only SS ex-PVPs are considered (Figure 2.9). A region of near fixation of the B73 haplotype in SS ex-PVPs was observed on chromosome 1 125-

150Mb. SS ex-PVPs across the three companies also shared a sharp drop in B73 haplotype sharing near the top of chromosome 3. Regions of differentiation between the companies were also observed such as those on chromosome 7 at 75-150Mb, chromosome 8 at 5-50Mb and chromosome 8 at 95-120Mb.

Discussion

Data integrity

Over the last several decades, seed sources for released, inbred lines have been maintained by repositories and breeding programs. Independent maintenance of seed sources through periodic seed increases can lead to phenotypic changes (Bogenschutz & Russell, 1986). Variation among seed sources within an inbred have been observed in multiple studies and attributed to factors such as residual heterozygosity, unintended introgression events, mutations, contamination and genetic drift (Gethi et al., 2002; Romay et al., 2013; Romero-Severson et al., 2001). The seed sources genotyped for a given inbred are assumed to be representative of the source used in breeding crosses. This may not always be the case and multiple sources could be independently used in breeding crosses (Haun et al., 2011; Liang & Schnable, 2016).

The SNP data present in HapMap 3.2.1 is a combination of multiple data sources (Bukowski et al., 2018). The data sources vary in the lines and seed sources sequenced, read length and quality and sequencing coverage depth. The variation in missing data and concordance among sample replicates can impact data integrity and the ability to accurately assess haplotype sharing. Collapsing of sample replicates decreased missing data for some inbreds in this study. Inbreds with source variation were identified through direct comparison of SNP calls among the replicates. Alternatively, a metric such as the Dice similarity index (Nei & Li, 1979) could be used. Five of the six inbreds, which had <97% match rate among

their replicates, showed widespread mismatches across the genome. Widespread mismatches can be indicative of contamination or mix-up at the seed source, sampling or DNA testing stage. Mismatches which are localized to genomic regions may be more indicative of residual heterozygosity that became fixed differently between or within a source or an unintended introgression event that may have become fixed in one source. Sample replicate analysis in conjunction with pedigree information was able to resolve half of the discordant replicate cases. Pedigree information can sometimes be inaccurate and incomplete (Messmer et al., 1993) so it may not be beneficial in all cases. When pedigree information are available, it can offer an advantage in analysis of sample integrity, regardless if one replicate or multiple replicates are present for a given inbred. In the presence of pedigree information, parent-progeny or triplet analysis can also be performed as a measure of the sample replicate integrity. Parent-progeny analysis was not emphasized in this study due to limited triplets available.

Analyses such as these are important so that researchers understand potential consequences of seed source variation and can ensure proper sources are selected for their study. When inconsistencies and unexpected relationships are observed, knowledge of the germplasm and breeding history can help to identify the source of the discordance (Lorenz & Hoegemeyer, 2013). Even still it may not be possible to determine the cause and not all seed sources for a given inbred are available for genetic testing, however, this information can better inform line selection and interpretation of results.

Haplotype sharing provides insight into maize industry breeding program history

The results presented here build on previous studies by providing high resolution haplotype data uncovering specific regions of haplotype sharing between ex-PVPs and founders and the similarities and differences across seed industry companies and heterotic

groups. Beckett et al. (2017) showed that commercial breeding efforts continue to drive divergence of heterotic groups. The haplotype sharing results support this conclusion with clear differences in founder haplotype sharing in the various heterotic groups. The proportion of haplotype groups shared between SS and NSS was similar to the proportion shared between NSS and Iodent (Table 2.1). This suggests that the NSS and Iodent heterotic groups are just as different from one another as NSS are from SS. Iodent is a strain of Reid Yellow Dent (Troyer, 1999) which was important in the development of the SS heterotic group. This relationship could explain the greater proportion of shared haplotype groups observed between Iodent and SS compared to Iodent and NSS.

Although the results are only reflective of the inbreds included in this study and several ex-PVPs were not included, the trends in haplotype sharing between ex-PVPs and founders are in line with expectations from pedigree-based studies (Mikel, 2011). A comparison of pedigree-based co-ancestry to % IBS is often used to identify unexpected relationships and confirm theoretical based inheritance (Bernardo, 1993; Inghelandt et al., 2010; Lübberstedt et al., 2000). This type of analysis can also be performed using the % IBS haplotype sharing values. To demonstrate the relationship between pedigree-based co-ancestry and the % IBS haplotype sharing values obtained in this study, 19 ex-PVPs with publicly available pedigree information were compared against the 12 key founders. Ex-PVPs were selected based on tracing back to at least 1 of the 12 founders where at least one founder was not a direct parent. The additive relationships based on pedigree information were obtained using the R package ‘AGHmatrix’. In cases of known ancestry, the % IBS haplotype sharing values are more closely aligned to the pedigree-based co-ancestry values than are SNP-based co-ancestry (IBS) values because the multi-SNP haplotypes are taking

advantage of the LD structure present (Figure A 2.2). PHG39, a DowDuPont ex-PVP, offers an example of an unexpected relationship with the founders B14 and B37. PHG39 has no known pedigree relationship to B14 or B37 but has 40.1% and 47.4% IBS haplotype sharing with these founders, respectively. Clustering of PHG39 with B14 and B37 has been observed in previous studies (Beckett et al., 2017; Kahler et al., 2010; Nelson et al., 2016), but a direct pedigree link remains unknown.

Analysis of additional ex-PVPs submitted for PVP from 1976 to 1992 would provide a more complete assessment of the haplotype sharing, but it is not expected the trends in haplotype sharing between ex-PVPs and founders would change drastically. It is also not expected that ex-PVPs registered after 1992 will stick to the same trends. Through each generation, recombination breaks up haplotypes which have not become fixed in a population. Further, if a particular segment of the germplasm has become less emphasized in industry breeding programs, it is likely that haplotype sharing with founders associated with this germplasm segment would reflect this.

Based on pedigree analysis by Mikel (2011), contributions of founders such as Mo17, B14 and B37 to ex-PVPs have decreased over time while contributions of founders such as 207 have increased. Although the counts are limited in the set of ex-PVPs used in this study, a breakout of the ex-PVPs by application years, 1976-1987 and 1988-1992, suggests a decrease in IBS sharing with the 12 founders from 77.4% to 74.4%. However, this decrease is primarily observed in DowDuPont ex-PVPs as Monsanto and Syngenta ex-PVPs experience an increase in haplotype sharing with 207 (Figure A 2.5) for ex-PVPs registered 1988-1992. As more ex-PVPs become available, the changes in these trends over time based on genotypic data may become more apparent.

These haplotype sharing results reveal extended haplotype sharing among some of the key founders and the regions at which this occurs. Although selection of the founders using previous literature was generally effective, resulting in much of the ex-PVP haplotype space IBS with at least one of the twelve founders, there is room for improvement. A632 was the 5th highest founder based on IBS haplotype sharing with all ex-PVPs, but the cumulative IBS haplotype sharing plot (Figure 2.5) shows minimal increase in IBS haplotypes accounted for by the addition of A632. The pedigree of A632 is [(Mt42*B14)B14(3)] (Gerdes et al., 1993) and significant haplotype sharing with B14 is observed. 76% of the A632 haplotype space is IBS with B14. The minimal increase that A632 adds to the cumulative distribution plot suggests that the A632 haplotypes shared with these ex-PVPs are primarily haplotypes that are IBS with B14. There remain haplotypes which could not be explained by these founders nor by the additional public lines included in the study. The additional public lines were selected based on being known ancestors to PVP lines. Given the proprietary nature of many of the pedigrees of PVPs, it is likely that there are additional public lines unaccounted for that may explain some of this additional haplotype space. With genotypic data now available across multiple studies covering a larger number of ex-PVPs, selection of key founders for ex-PVP germplasm could be improved by using a marker-based probability of gene origin approach described by Technow et al. (2014).

The haplotypes from this study not only provide insight into these ex-PVP lines registered 1976-1992, but the haplotype data can be used to infer haplotypes of parents and progeny inbreds which were not included either due to lack of genotypic data or the proprietary nature of the lines. In a two-parent cross where haplotype data are available for both a progeny inbred and one of its two parents, any haplotypes in the progeny inbred that

are not IBS with the parent haplotypes can be assumed to have originated in the non-haplotyped parent. Similarly, these haplotypes can provide insight into PVPs which have yet to expire. If both parents of a given PVP have been haplotyped, the haplotype structure can be inferred based on the combination of these parents, e.g. haplotypes that are IBS between the parents will appear in the progeny inbred.

Application of haplotype sharing analysis to breeding

The results from this study demonstrate how haplotype analysis can be used to understand germplasm diversity. Haplotype-based analyses provide advantages over single-marker based approaches for inference of population structure (Gattepaille & Jakobsson, 2012; Haasl & Payseur, 2011). Haplotypes can provide additional information compared to SNPs as they combine allele information across multiple SNPs. When SNPs are dense enough to capture the LD that exists in a population, haplotypes formed by the SNPs can provide more power to analyses. Previous studies have successfully utilized haplotypes in assessment of population structure in species such as barley (Fang et al., 2014) and humans (Lawson et al., 2012; Ralph & Coop, 2013). Haplotype sharing has been used as a measure of relatedness and can reveal regions of selection and variation in diversity present within germplasm (Fang et al., 2014; Hufford et al., 2013; Poets et al., 2016). Low regions of diversity may highlight opportunities for increasing diversity within a breeding program. Alternatively, low diversity regions may be present due to selective sweeps (Jiao et al., 2012). Diversity trends through haplotype structure can be monitored over time within a breeding program to ensure haplotypes are not fixed unintentionally. Further, use of haplotypes in genomic predictions can increase prediction accuracy over single-marker based approaches but are influenced by haplotype length and the trait of interest (Ferdosi et al., 2016). Schrag et al. (2007) demonstrated use of haplotype-based prediction in a maize hybrid

breeding program. Recently, Jiang et al. (2018) showed that modeling haplotype effects can increase prediction accuracy and capture epistatic interactions in mouse, rice and maize.

Knowledge of haplotype structure in the breeding germplasm can help inform future sequencing and genotyping strategy (Ros-Freixedes et al., 2017). High-resolution haplotype structure can be assessed from deep sequencing of a small set of key inbreds representative of the population or breeding program. A reduced set of loci that tag each haplotype (tag SNPs) can be utilized to more effectively and cost efficiently genotype samples for downstream analyses. Tag SNPs have been utilized in previous studies resulting in increased mapping efficiency (Kebede et al., 2016; Y. Lu et al., 2010). Similarly, samples can be sequenced at low coverage and missing data imputed using the high-resolution haplotypes, thereby reducing sequencing cost. In this study, the public lines, and in particular the 12 key founders, had much lower missing SNP data while the ex-PVP lines generally had higher missing data. The resulting high proportion of the ex-PVP genomes which were able to be assigned to haplotype groups highlights the effectiveness of a sequencing strategy in which the key inbreds receive deep sequencing and other individuals receive lower coverage sequencing.

One limitation in this study is the use of a single reference genome. The haplotypes observed are a direct reflection of sequencing reads which were only aligned to the B73 AGPv4 reference genome. The B73 reference genome does not capture the full extent of genome variation within the species (Brunner et al., 2005; Lu et al., 2015). High levels of copy number variants and presence/absence variants have been observed in multiple studies (Hirsch et al., 2016; Lai et al., 2010; Springer et al., 2009). Because of this limitation, there are certainly haplotypes that exist in these ex-PVPs that are not captured in this study.

Genome graphs are being explored as a method to build pan-genomes (Paten et al., 2017) to account for the variation within a species. In maize and other species, a Practical Haplotype Graph is being implemented to capture haplotypes across a pan-genome framework (Johnson et al., 2018).

Accumulation of deleterious alleles in maize has been suggested to be primarily driven by the effects of the domestication bottleneck (Wang et al., 2017). Ramu et al. (2017) showed haplotypes containing fewer deleterious alleles have been favored during selection in cassava, but that drift has increased fixation of deleterious alleles. Deleterious alleles have been shown to affect expression in maize (Kremling et al., 2018) and have been proposed as a major driver in hybrid vigor through incomplete dominance (Yang et al., 2017). Deleterious alleles can be identified through genomic evolutionary rate profiling (GERP; Davydov et al., 2010). Complementation of deleterious alleles is thought to be involved in heterosis (Lai et al., 2010). The 11.3 million loci in this study were divided into 5,694 haplotype blocks each containing ~2,000 SNPs. These data present an opportunity to look at haplotype-specific accumulation of deleterious alleles with high-density data and the influence of long-range haplotype sharing on interpretation of GERP scores. If differential accumulation of deleterious alleles in the haplotypes is observed, these data could be compared with the known heterotic group structure to see if complementation is significant between heterotic groups as compared to random.

In summary, this study provides increased resolution to the understanding of genetic relationships between maize ex-PVPs and key founders. Trends in haplotype sharing between the 12 key founders and ex-PVPs confirm trends from pedigree-based estimates. A majority of the haplotype space in the ex-PVPs analyzed can be accounted for by just a few key

founders. Comparison of haplotype structure across seed industry companies and heterotic groups reveals broad-scale patterns of haplotype sharing and regional differences. Relating the haplotypes present in ex-PVPs to key founders provides an intuitive way to understand the breeding history of industry germplasm and can support future studies of trait-associated haplotypes, facilitate selection of ex-PVP lines for use in breeding programs, studies of diversity and heterosis and inference of haplotypes in PVPs not yet expired.

Figures

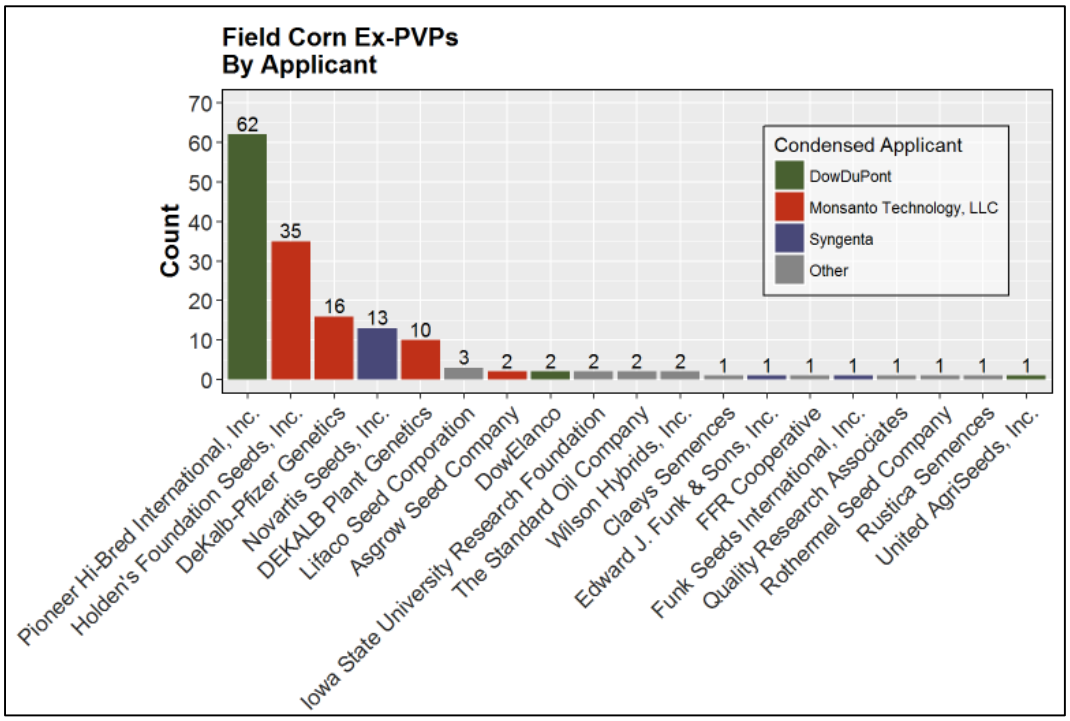

Figure 2.1. Breakout of the 157 field corn ex-PVPs by applicant. Applicants are colored based on the current company when mergers and acquisitions are taken into consideration.

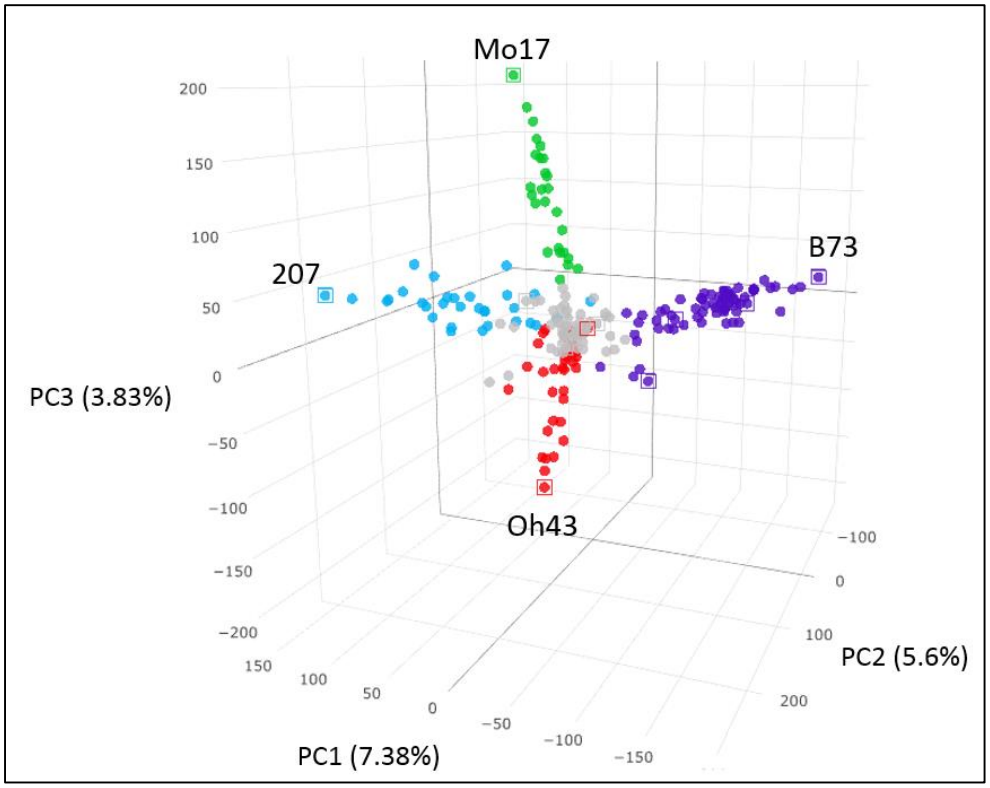

Figure 2.2. 3D principal components plot for 157 ex-PVPs and 55 public lines. Points are colored based on heterotic group assignment (SS = purple, NSS = green, Iodent = blue, Oh43-type = red) and the key founders labeled for each heterotic group. Squares encompass the points for each of the 12 founders.

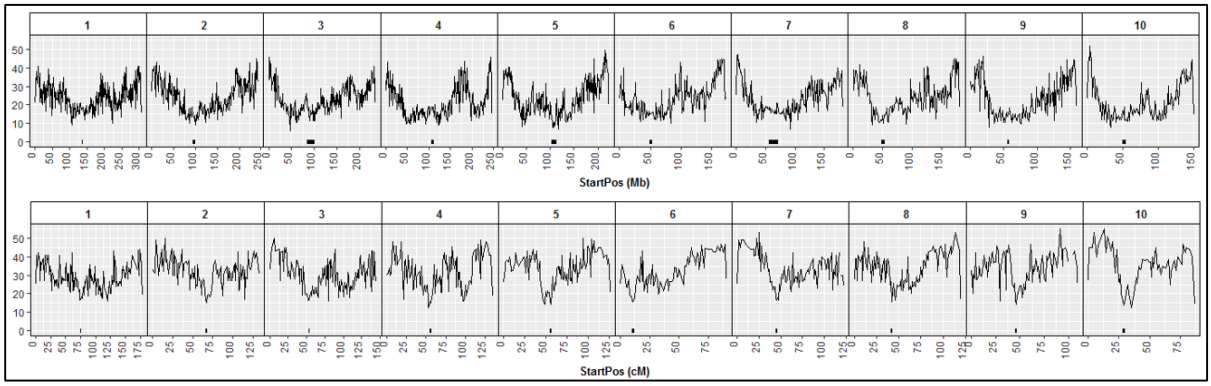

Figure 2.3. Average number of haplotype groups by physical (Mb) and genetic (cM) bins. The black line represents the average number of haplotype groups present within a given bin across the genome. The top panel shows averages by 1Mb bin and the bottom panel shows averages by 1cM bin. Approximate centromere positions are represented by the black rectangles along the x-axis.

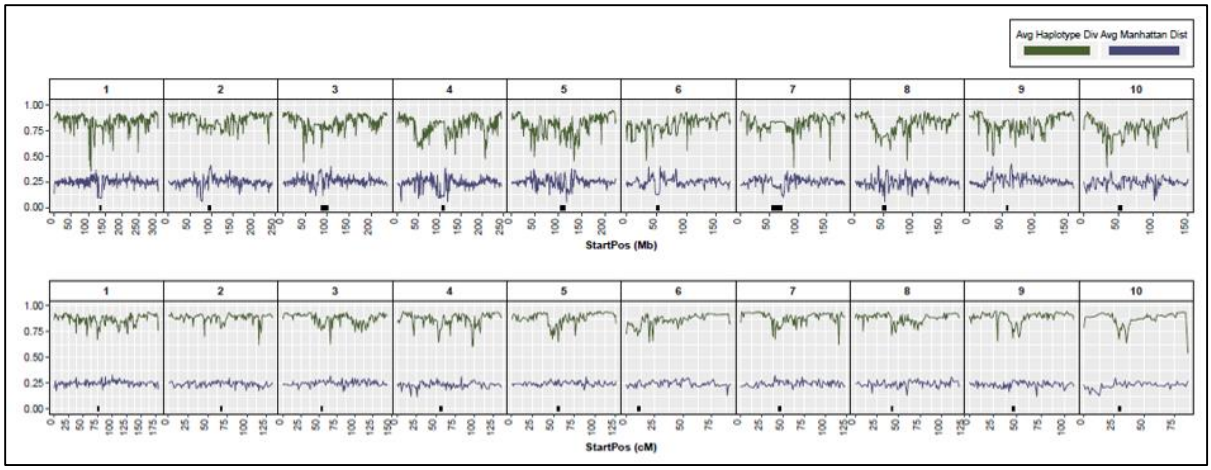

Figure 2.4. Haplotype diversity (*H*) and Manhattan distance among consensus haplotypes. The green line represents average *H* in 1Mb (top panel) and 1cM (bottom panel) bins. The blue line represents average Manhattan distance among consensus haplotypes in 1Mb (top panel) and 1cM (bottom panel) bins. Approximate centromere positions are represented by the black rectangles along the x-axis.

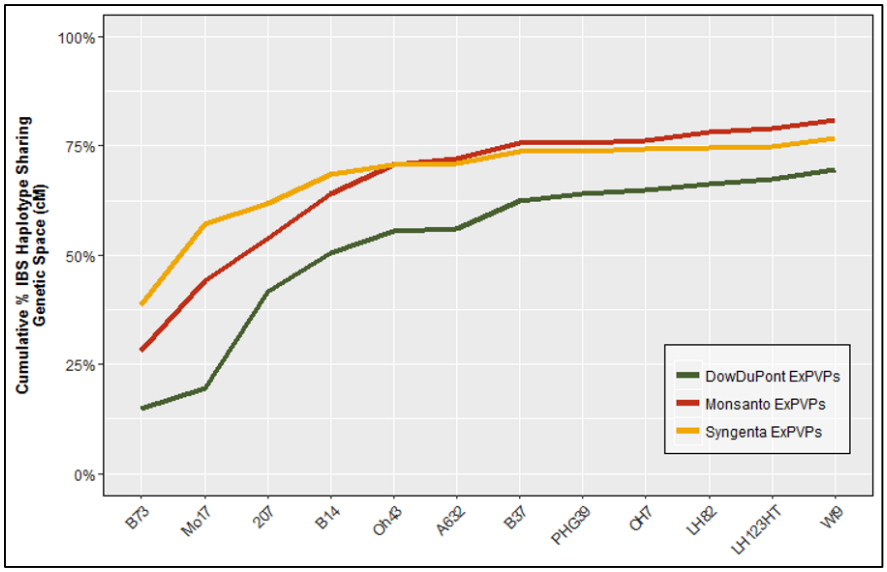

Figure 2.5. Cumulative haplotype sharing between ex-PVPs and the 12 key founders. Founder names are shown on the x-axis and % IBS haplotype sharing values on the y-axis. The order of the founder names is somewhat arbitrary, however, founders which displayed high levels of haplotype sharing with ex-PVPs or are key founders for heterotic groups were placed on the left. This is a cumulative plot and should be read from left to right. Starting with B73, the y-axis shows the % IBS haplotype sharing with each group of ex-PVPs. Moving to Mo17, the y-axis values increase based on the % IBS haplotype sharing which is not accounted for my B73. In other words, if we look for haplotypes which are shared between ex-PVPs and either B73 or Mo17, this is the % IBS haplotype sharing we observe. Moving to 207, the % IBS haplotype sharing increases based on haplotype sharing which is not accounted for by B73 or Mo17. This continues cumulatively from left to right.

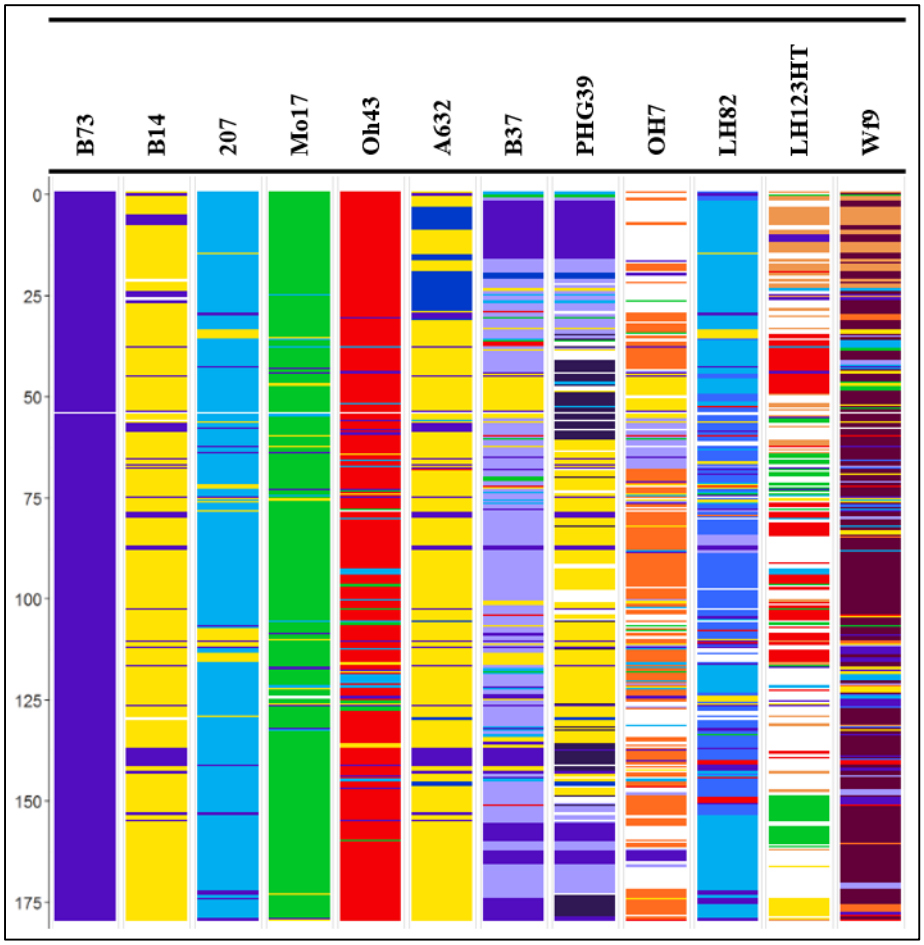

Figure 2.6. Haplotype assignments for the 12 key founders on chromosome 1. Genetic positions are shown on the y-axis and the x-axis contains each of the 12 founders. Haplotypes are colored in a hierarchical manner from left to right. First, B73 receives a single color (purple). The next founder, B14 displays purple in regions it shares haplotypes with B73 and a new color (yellow) at regions it differs. At the third founder, 207, regions where 207 is not IBS with B73 nor B14 receive a blue color. This continues left to right through the remaining founders. Regions of the chromosome that did not receive a haplotype assignment and are thus missing are colored white. Regions where founders have identical colors, such as the region near the top of chromosome 1 for B73, B37 and PHG39, are regions of IBS haplotype sharing.

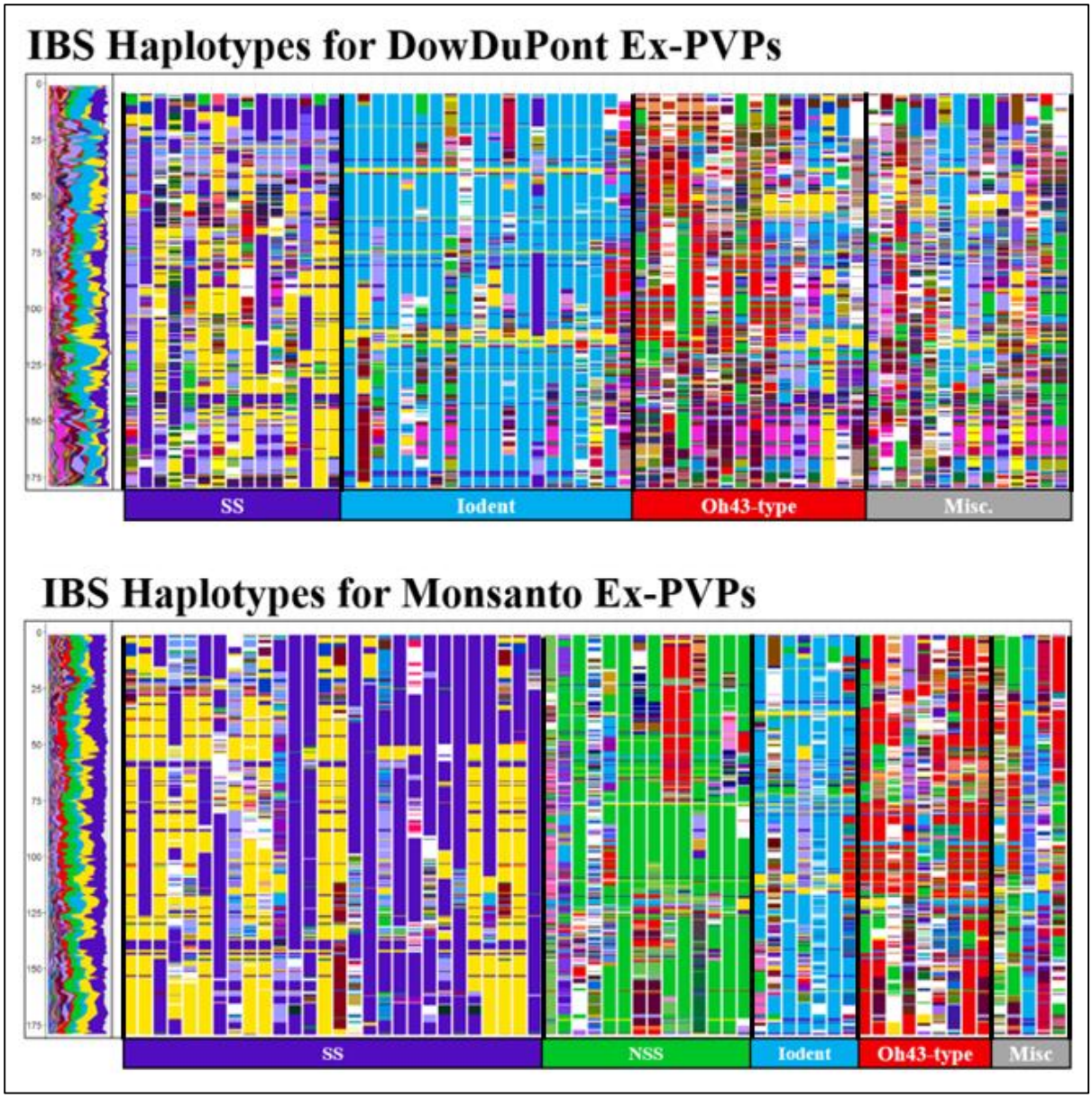

Figure 2.7. IBS haplotype assignments on chromosome 1 for DowDuPont (top panel) and Monsanto (bottom panel) ex-PVPs. Inbreds are ordered left to right first based on heterotic group assignment and then alphabetic based on the sample name. Genetic positions are displayed on the y-axis. Shared colors across inbreds represent regions that are IBS between those inbreds. The bacon-like view on the left shows a condensed form or composite view of the haplotypes across the given set of lines.

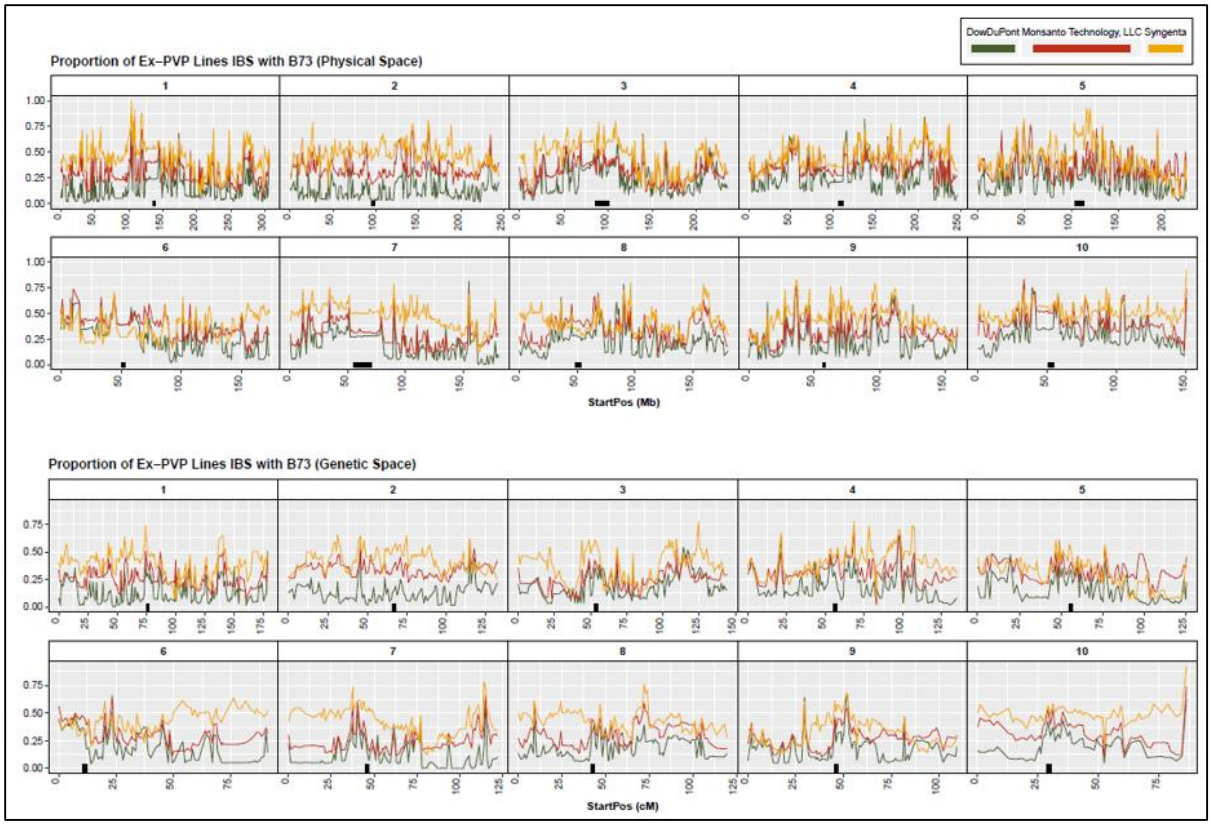

Figure 2.8. Proportion of IBS haplotype sharing between ex-PVPs and B73. The top panel displays average IBS haplotype sharing in 1Mb bins. The bottom panel displays average haplotype sharing in 1cM bins.

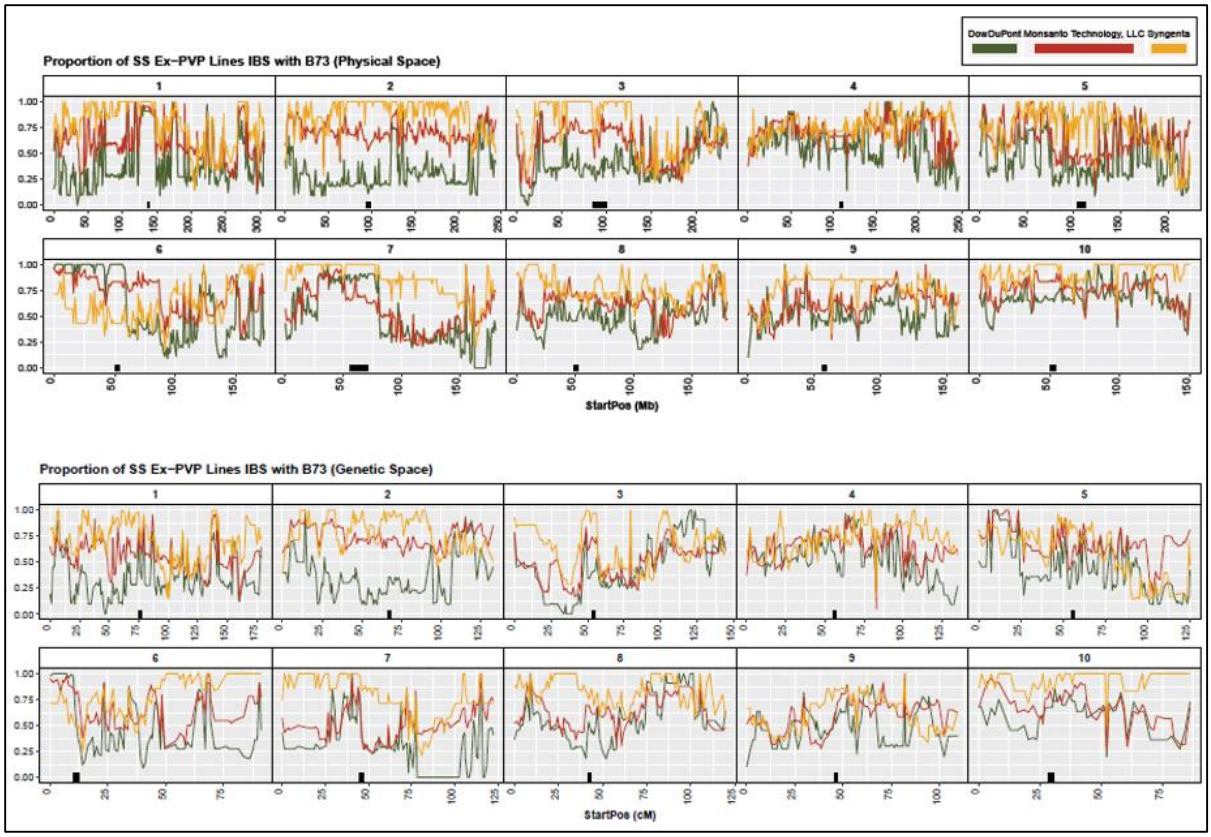

Figure 2.9. Proportion of IBS haplotype sharing between SS ex-PVPs and B73. The top panel displays average IBS haplotype sharing in 1Mb bins. The bottom panel displays IBS haplotype sharing in 1cM bins.

Tables

Table 2.1. Shared haplotype groups between heterotic groups. The numbers in the upper triangle represent the number of haplotype groups in which at least 1 individual from each of the heterotic groups in the pair were placed into the same haplotype group. The lower triangle (gray shading) represents the proportion of haplotype groups present among the pair of heterotic groups which were shared.

Heterotic Group	SS	NSS	Iodent	Oh43-type
SS	45,199	14,389	19,127	20,831
NSS	21.5%	36,207	13,920	20,759
Iodent	31.8%	24.7%	34,087	20,214
Oh43-type	27.8%	31.5%	31.4%	50,531

Table 2.2. % IBS haplotype sharing between ex-PVPs and the 12 founders. Counts for All ExPVPs, DowDuPont Ex-PVPs, Monsanto Ex-PVPs and Syngenta Ex-PVPs are 157, 65, 63 and 15, respectively. The % haplotype sharing is derived as the proportion of haplotype space in an ex-PVP that is shared with the given founder. The % haplotype sharing values are then averaged among each respective group of ex-PVPs.

Founder	All ExPVPs	DowDuPont	Monsanto	Syngenta
207	19.03%	25.37%	13.37%	8.36%
A632	14.53%	13.22%	17.53%	14.23%
B14	16.55%	14.73%	20.00%	17.31%
B37	13.69%	15.28%	13.69%	13.15%
B73	22.02%	14.52%	27.24%	36.81%
LH123HT	9.16%	8.06%	10.56%	7.09%
LH82	9.27%	10.01%	9.17%	5.57%
Mo17	14.84%	7.37%	19.81%	23.05%
Oh43	9.28%	7.70%	10.15%	5.42%
OH7	4.97%	5.03%	5.24%	5.04%
PHG39	14.00%	16.15%	13.96%	11.83%
Wf9	7.62%	7.89%	7.15%	6.90%

Data availability

The dataset containing the haplotype group assignment results is available at

https://github.com/scoffman/maize_expvp_haplotypes

Author contribution statement

SMC designed the study, designed and ran the analyses, interpreted the results and wrote the manuscript. MBH, CMA and TH assisted with the study design, interpretation of results and provided critical feedback during manuscript preparation. TH supervised the study.

Acknowledgements

The author SMC wishes to thank Justin Gerke and Dean Podlich for helpful discussions in preparation of this manuscript.

Conflict of interest

The author SMC is employed by DowDuPont. The funder provided support in the form of salary and graduate program support for author SMC, but did not have any additional role in the study design, data collection and analysis, decision to publish or preparation of the manuscript. The specific roles of this author are articulated in the author contribution statement. The authors declare that they have no conflict of interest.

Ethical standards

The authors declare that ethical standards are met, and all the experiments comply with the current laws of the country in which they were performed.

References

Beckett, T. J., Morales, A. J., Koehler, K. L., & Rocheford, T. R. (2017). Genetic relatedness of previously Plant-Variety-Protected commercial maize inbreds. *PLoS ONE*, *12*(12). https://doi.org/10.1371/journal.pone.0189277

Bernardo, R. (1993). Estimation of coefficient of coancestry using molecular markers in maize. *Theoretical and Applied Genetics*, *85*(8), 1055–1062. https://doi.org/10.1007/BF00215047

Bernardo, R., Romero-Severson, J., Ziegle, J., Hauser, J., Joe, L., Hookstra, G., & Doerge, R. W. (2000). Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP, and SSR data. *Theoretical and Applied Genetics*, *100*(3–4), 552–556. https://doi.org/10.1007/s001220050072

Bogenschutz, T. G., & Russell, W. A. (1986). An evaluation for genetic variation within maize inbred lines maintained by sib-mating and self-pollination. *Euphytica*, *35*(2), 403–412. https://doi.org/10.1007/BF00021848

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Brunner, S., Fengler, K., Morgante, M., Tingey, S., & Rafalski, A. (2005). Evolution of DNA Sequence Nonhomologies among Maize Inbreds. *The Plant Cell*, *17*(2), 343–360. https://doi.org/10.1105/tpc.104.025627

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., … Xu, Y. (2018). Construction of the third-generation Zea mays haplotype map. *GigaScience*, *7*(4). https://doi.org/10.1093/gigascience/gix134

Davydov, E. V., Goode, D. L., Sirota, M., Cooper, G. M., Sidow, A., & Batzoglou, S. (2010). Identifying a High Fraction of the Human Genome to be under Selective Constraint Using GERP++. *PLOS Computational Biology*, *6*(12), e1001025. https://doi.org/10.1371/journal.pcbi.1001025

Dell'Acqua, M., Gatti, D. M., Pea, G., Cattonaro, F., Coppens, F., Magris, G., … Pè, M. E. (2015). Genetic properties of the MAGIC maize population: a new platform for high definition QTL mapping in Zea mays. *Genome Biology*, *16*, 167. https://doi.org/10.1186/s13059-015-0716-z

East, E. M. (1908). Inbreeding in corn. *Connecticut Agric. Exp. Stn. Rep.*, *1907*, 419–428.

Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J., & Morrell, P. L. (2014). Two genomic regions contribute disproportionately to geographic differentiation in wild barley. *G3: Genes, Genomes, Genetics*, *4*(7), 1193–1203. https://doi.org/10.1534/g3.114.010561

Ferdosi, M. H., Henshall, J., & Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. *Genetics Selection Evolution*, *48*, 75. https://doi.org/10.1186/s12711-016-0253-6

Ganal, M. W., Durstewitz, G., Polley, A., Bérard, A., Buckler, E. S., Charcosset, A., … Falque, M. (2011). A Large Maize (Zea mays L.) SNP Genotyping Array: Development and Germplasm Genotyping, and Genetic Mapping to Compare with the B73 Reference Genome. *PLOS ONE*, *6*(12), e28334. https://doi.org/10.1371/journal.pone.0028334

Gattepaille, L. M., & Jakobsson, M. (2012). Combining Markers into Haplotypes Can Improve Population Structure Inference. *Genetics*, *190*(1), 159–174. https://doi.org/10.1534/genetics.111.131136

Gerdes, J. T., Behr, C. F., Coors, J. G., Tracy, W. F., Gerdes, J. T., Behr, C. F., … Tracy, W. F. (1993). Field Corn Inbred Lines. In *ACSESS publications*. Crop Science Society of America. https://doi.org/10.2135/1993.compilationofnorthamerican.c1

Gerke, J. P., Edwards, J. W., Guill, K. E., Ross-Ibarra, J., & McMullen, M. D. (2015). The Genomic Impacts of Drift and Selection for Hybrid Performance in Maize. *Genetics*, *201*(3), 1201–1211. https://doi.org/10.1534/genetics.115.182410

Gethi, J. G., Labate, J. A., Lamkey, K. R., Smith, M. E., & Kresovich, S. (2002). SSR Variation in Important U.S. Maize Inbred Lines. *Crop Science*, *42*(3), 951–957. https://doi.org/10.2135/cropsci2002.9510

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. https://doi.org/10.1038/hdy.2010.21

Haun, W. J., Hyten, D. L., Xu, W. W., Gerhardt, D. J., Albert, T. J., Richmond, T., … Stupar, R. M. (2011). The Composition and Origins of Genomic Variation among Individuals of the Soybean Reference Cultivar Williams 82. *Plant Physiology*, *155*(2), 645–655. https://doi.org/10.1104/pp.110.166736

Heerwaarden, J. van, Hufford, M. B., & Ross-Ibarra, J. (2012). Historical genomics of North American maize. *Proceedings of the National Academy of Sciences of the United States of America*, *109*(31), 12420–12425. https://doi.org/10.1073/pnas.1209275109

Hirsch, C., Hirsch, C. D., Brohammer, A. B., Bowman, M. J., Soifer, I., Barad, O., … Mikel, M. A. (2016). Draft Assembly of Elite Inbred Line PH207 Provides Insights into Genomic and Transcriptome Diversity in Maize, tpc.00353.2016. https://doi.org/10.1105/tpc.16.00353

Hufford, M. B., Lubinksy, P., Pyhäjärvi, T., Devengenzo, M. T., Ellstrand, N. C., & Ross-Ibarra, J. (2013). The Genomic Signature of Crop-Wild Introgression in Maize. *PLOS Genetics*, *9*(5), e1003477. https://doi.org/10.1371/journal.pgen.1003477

Hufford, M. B., Xu, X., van Heerwaarden, J., Pyhäjärvi, T., Chia, J.-M., Cartwright, R. A., … Ross-Ibarra, J. (2012). Comparative population genomics of maize domestication and improvement. *Nature Genetics*, *44*(7), 808–811. https://doi.org/10.1038/ng.2309

Inghelandt, D. V., Melchinger, A. E., Lebreton, C., & Stich, B. (2010). Population structure and genetic diversity in a commercial maize breeding program assessed with SSR and SNP markers. *Theoretical and Applied Genetics*, *120*(7), 1289–1299. https://doi.org/10.1007/s00122-009-1256-2

Jiang, Y., Schmidt, R. H., & Reif, J. C. (2018). Haplotype-Based Genome-Wide Prediction Models Exploit Local Epistatic Interactions Among Markers. *G3: Genes, Genomes, Genetics*, g3.300548.2017. https://doi.org/10.1534/g3.117.300548

Jiao, Y., Zhao, H., Ren, L., Song, W., Zeng, B., Guo, J., … Lai, J. (2012). Genome-wide genetic changes during modern breeding of maize. *Nature Genetics*, *44*(7), 812–815. https://doi.org/10.1038/ng.2312

Johnson, L. C., Bradbury, P., Casstevens, T., Ilut, D., Miller, Z., Punna, R., … Buckler, E. (2018). *A Practical Haplotype Graph for Determining Genomic Sequence*. Poster presented at the Plant & Animal Genome Conference XXVI, San Diego, CA.

Kahler, A. L., Kahler, J. L., Thompson, S. A., Ferriss, R. S., Jones, E. S., Nelson, B. K., … Smith, S. (2010). North American Study on Essential Derivation in Maize: II. Selection and Evaluation of a Panel of Simple Sequence Repeat Loci. *Crop Science*, *50*(2), 486. https://doi.org/10.2135/cropsci2009.03.0121

Kaufman, L., & Rousseeuw, P. J. (2008). Partitioning Around Medoids (Program PAM). In *Finding Groups in Data: An Introduction to Cluster Analysis*. Hoboken, NJ, USA: John Wiley & Sons, Inc. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470316801.ch2/summary

Kebede, A. Z., Woldemariam, T., Reid, L. M., & Harris, L. J. (2016). Quantitative trait loci mapping for Gibberella ear rot resistance and associated agronomic traits using genotyping-by-sequencing in maize. *Theoretical and Applied Genetics*, *129*(1), 17–29. https://doi.org/10.1007/s00122-015-2600-3

Kremling, K. A. G., Chen, S.-Y., Su, M.-H., Lepak, N. K., Romay, M. C., Swarts, K. L., … Buckler, E. S. (2018). Dysregulation of expression correlates with rare-allele burden and fitness loss in maize. *Nature*. https://doi.org/10.1038/nature25966

Lai, J., Li, R., Xu, X., Jin, W., Xu, M., Zhao, H., … Wang, J. (2010). Genome-wide patterns of genetic variation among elite maize inbred lines. *Nature Genetics*, *42*(11), 1027–1030. https://doi.org/10.1038/ng.684

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Liang, Z., & Schnable, J. C. (2016). RNA-Seq Based Analysis of Population Structure within the Maize Inbred B73. *PLOS ONE*, *11*(6), e0157942. https://doi.org/10.1371/journal.pone.0157942

Lorenz, A., & Hoegemeyer, T. (2013). The phylogenetic relationships of US maize germplasm. *Nature Genetics*, *45*(8), 844–845. https://doi.org/10.1038/ng.2697

Lu, F., Romay, M. C., Glaubitz, J. C., Bradbury, P. J., Elshire, R. J., Wang, T., … Buckler, E. S. (2015). High-resolution genetic mapping of maize pan-genome sequence anchors. *Nature Communications*, *6*, ncomms7914. https://doi.org/10.1038/ncomms7914

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., … Xu, Y. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences*, *107*(45), 19585–19590. https://doi.org/10.1073/pnas.1006105107

Lübberstedt, T., Melchinger, A. E., Dußle, C., Vuylsteke, M., & Kuiper, M. (2000). Relationships among Early European Maize Inbreds: IV. Genetic Diversity Revealed with AFLP Markers and Comparison with RFLP, RAPD, and Pedigree Data. *Crop Science*, *40*(3), 783–791. https://doi.org/10.2135/cropsci2000.403783x

Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M., & Hornik, K. (2018). *cluster: Cluster Analysis Basics and Extensions*.

Mangelsdorf, P. C. (1975). Donald Forsha Jones (1890-1963). In *Biographical Memoirs* (Vol. 46, pp. 135–156). National Academy of Science. Retrieved from https://doi.org/10.17226/569

Messmer, M. M., Melchinger, A. E., Herrmann, R. G., & Boppenmaier, J. (1993). Relationships among Early European Maize Inbreds: II. Comparison of Pedigree and RFLP Data. *Crop Science*, *33*(5), 944–950. https://doi.org/10.2135/cropsci1993.0011183X003300050014x

Mikel, M. A. (2008). Genetic Diversity and Improvement of Contemporary Proprietary North American Dent Corn. *Crop Science*, *48*(5), 1686. https://doi.org/10.2135/cropsci2008.01.0039

Mikel, M. A. (2011). Genetic Composition of Contemporary U.S. Commercial Dent Corn Germplasm. *Crop Science*, *51*(2), 592–599. https://doi.org/10.2135/cropsci2010.06.0332

Mikel, M. A., & Dudley, J. W. (2006). Evolution of North American Dent Corn from Public to Proprietary Germplasm. *Crop Science*, *46*(3), 1193. https://doi.org/10.2135/cropsci2005.10-0371

Nei, M, & Li, W. H. (1979). Mathematical model for studying genetic variation in terms of restriction endonucleases. *Proceedings of the National Academy of Sciences of the United States of America*, *76*(10), 5269–5273.

Nei, Masatoshi. (1987). *Molecular Evolutionary Genetics*. Columbia University Press.

Nelson, Paul T., Coles, N. D., Holland, J. B., Bubeck, D. M., Smith, S., & Goodman, M. M. (2008). Molecular Characterization of Maize Inbreds with Expired U.S. Plant Variety Protection. *Crop Science*, *48*(5), 1673–1685. https://doi.org/10.2135/cropsci2008.02.0092

Nelson, P.T., Krakowsky, M. D., Coles, N. D., Holland, J. B., Bubeck, D. M., Smith, J. S. C., & Goodman, M. M. (2016). Genetic Characterization of the North Carolina State University Maize Lines. *Crop Science*, *56*(1), 259. https://doi.org/10.2135/cropsci2015.09.0532

Paten, B., Novak, A. M., Eizenga, J. M., & Garrison, E. (2017). Genome graphs and the evolution of genome inference. *Genome Research*, gr.214155.116. https://doi.org/10.1101/gr.214155.116

Poets, A. M., Mohammadi, M., Seth, K., Wang, H., Kono, T. J. Y., Fang, Z., … Morrell, P. L. (2016). The Effects of Both Recent and Long-Term Selection and Genetic Drift Are Readily Evident in North American Barley Breeding Populations. *G3: Genes, Genomes, Genetics*, *6*(3), 609–622. https://doi.org/10.1534/g3.115.024349

R Core Team. (2016). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ralph, P., & Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. *PLOS Biology*, *11*(5), e1001555. https://doi.org/10.1371/journal.pbio.1001555

Ramu, P., Esuma, W., Kawuki, R., Rabbi, I. Y., Egesi, C., Bredeson, J. V., … Lu, F. (2017). Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. *Nature Genetics*, *49*(6), 959–963. https://doi.org/10.1038/ng.3845

Romay, M. C., Millard, M. J., Glaubitz, J. C., Peiffer, J. A., Swarts, K. L., Casstevens, T. M., … Gardner, C. A. (2013). Comprehensive genotyping of the USA national maize inbred seed bank. *Genome Biology*, *14*, R55. https://doi.org/10.1186/gb-2013-14-6-r55

Romero-Severson, J., Smith, J. S. C., Ziegle, J., Hauser, J., Joe, L., & Hookstra, G. (2001). Pedigree analysis and haplotype sharing within diverse groups of Zea mays L. inbreds. *Theoretical and Applied Genetics*, *103*(4), 567–574. https://doi.org/10.1007/PL00002911

Ros-Freixedes, R., Gonen, S., Gorjanc, G., & Hickey, J. M. (2017). A method for allocating low-coverage sequencing resources by targeting haplotypes rather than individuals. *Genetics Selection Evolution*, *49*, 78. https://doi.org/10.1186/s12711-017-0353-y

Schrag, T. A., Maurer, H. P., Melchinger, A. E., Piepho, H.-P., Peleman, J., & Frisch, M. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theoretical and Applied Genetics*, *114*(8), 1345–1355. https://doi.org/10.1007/s00122-007-0521-5

Shull, G. H. (1908). The Composition of a Field of Maize. *Journal of Heredity*, *os-4*(1), 296–301. https://doi.org/10.1093/jhered/os-4.1.296

Smith, J. S. C., Duvick, D. N., Smith, O. S., Grunst, A., & Wall, S. J. (1999). Effect of Hybrid Breeding on Genetic Diversity in Maize. In *The Genetics and Exploitation of Heterosis in Crops* (James G. Coors and Shivaji Pandey (ed.), Vol. acsesspublicati, pp. 119–126). Madison, Wisconsin: Crop Science Society of America. Retrieved from https://dl-sciencesocieties-org.proxy.lib.iastate.edu/publications/books/articles/acsesspublicati/thegeneticsande/119

Smith, S. (2007). Pedigree Background Changes in U.S. Hybrid Maize between 1980 and 2004. *Crop Science*, *47*(5), 1914–1926. https://doi.org/10.2135/cropsci2006.12.0763

Springer, N. M., Ying, K., Fu, Y., Ji, T., Yeh, C.-T., Jia, Y., … Schnable, P. S. (2009). Maize Inbreds Exhibit High Levels of Copy Number Variation (CNV) and Presence/Absence Variation (PAV) in Genome Content. *PLOS Genetics*, *5*(11), e1000734. https://doi.org/10.1371/journal.pgen.1000734

Swarts, K., Li, H., Navarro, R., Alberto, J., An, D., Romay, M. C., … Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome*, *7*(3). https://doi.org/10.3835/plantgenome2014.05.0023

Technow, F., Schrag, T. A., Schipprack, W., & Melchinger, A. E. (2014). Identification of key ancestors of modern germplasm in a breeding program of maize. *Theoretical and Applied Genetics*, *127*(12), 2545–2553. https://doi.org/10.1007/s00122-014-2396-6

Tracy, W. F., & Chandler, M. A. (2006). The Historical and Biological Basis of the Concept of Heterotic Patterns in Corn Belt Dent Maize. In K. R. Lamkey & M. Lee (Eds.), *Plant Breeding: The Arnel R. Hallauer International Symposium* (pp. 219–233). Blackwell Publishing. Retrieved from http://onlinelibrary.wiley.com/doi/10.1002/9780470752708.ch16/summary

Troyer, A. F. (1999). Background of U.S. Hybrid Corn. *Crop Science*, *39*(3), 601–626. https://doi.org/10.2135/cropsci1999.0011183X003900020001x

Wang, L., Beissinger, T. M., Lorant, A., Ross-Ibarra, C., Ross-Ibarra, J., & Hufford, M. B. (2017). The interplay of demography and selection during maize domestication and expansion. *Genome Biology*, *18*. https://doi.org/10.1186/s13059-017-1346-4

Wu, X., Li, Y., Fu, J., Li, X., Li, C., Zhang, D., … Wang, T. (2016). Exploring Identity-By-Descent Segments and Putative Functions Using Different Foundation Parents in Maize. *PLOS ONE*, *11*(12), e0168374. https://doi.org/10.1371/journal.pone.0168374

Yang, J., Mezmouk, S., Baumgarten, A., Buckler, E. S., Guill, K. E., McMullen, M. D., … Ross-Ibarra, J. (2017). Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. *PLoS Genetics*, *13*(9). https://doi.org/10.1371/journal.pgen.1007019

Zeileis, A., & Grothendieck, G. (2005). zoo: S3 Infrastructure for Regular and Irregular Time Series. *Journal of Statistical Software, Articles*, *14*(6), 1--27. https://doi.org/10.18637/jss.v014.i06

Zhou, X., Carter, T. E., Cui, Z., Miyazaki, S., & Burton, J. W. (2000). Genetic Base of Japanese Soybean Cultivars Released during 1950 to 1988. *Crop Science*, *40*(6), 1794–1802. https://doi.org/10.2135/cropsci2000.4061794x

Appendix

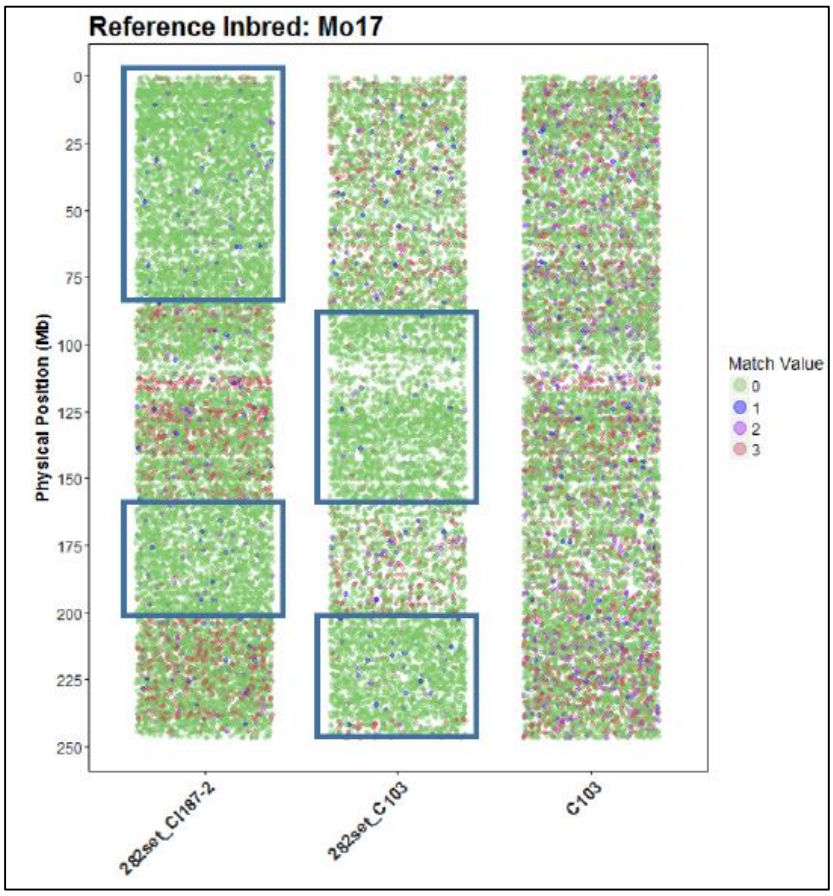

Figure A 2.1. Comparison of SNP calls on chromosome 4 for replicates '282set_CI187-2', '282set_C103' and 'C103' against a Mo17 replicate. C103 and CI187-2 are the inbred parents of Mo17. Points are colored based on match type. Green are SNP calls that match between the given replicate and Mo17. Blue are calls where Mo17 has a het call and the given rep has a homozygous call. Purple are calls where Mo17 has a homozygous call and the given rep has a het call. Red are calls where the replicates have differing homozygous calls. The visual comparison of each of the C103 replicates to the Mo17 replicates revealed the '282set_C103' replicate had more contiguous blocks of SNP calls matching to Mo17 (blue rectangles) suggestive of a possible inheritance pattern. CI187-2 and the 2 C103 replicates were then compared together against the Mo17 replicates. In line with inheritance assumptions, when Mo17 was compared to CI187-2 and '282set_C103', contiguous blocks that alternated between matching CI187-2 and matching '282set_C103' were observed The '282set_C103' replicate was thus selected and the 'C103' replicate discarded.

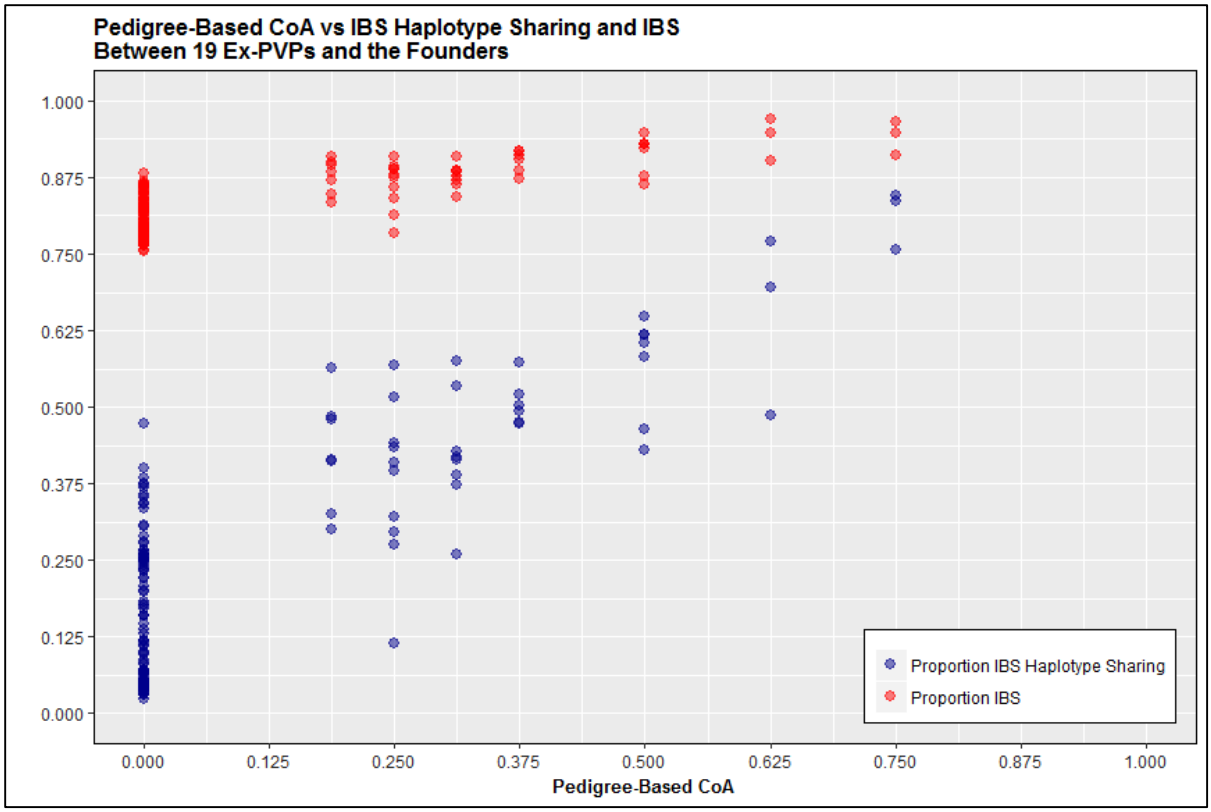

Figure A 2.2. Comparison of the pedigree-based co-ancestry values against IBS similarity and IBS haplotype sharing values. Pedigree-based co-ancestry values are on the x-axis and % pairwise IBS (red) and pairwise % haplotype sharing (blue) are on the y-axis. Comparisons are between 19 ex-PVPs and at least 1 of the 12 founders which are present in the pedigree of the ex-PVP.

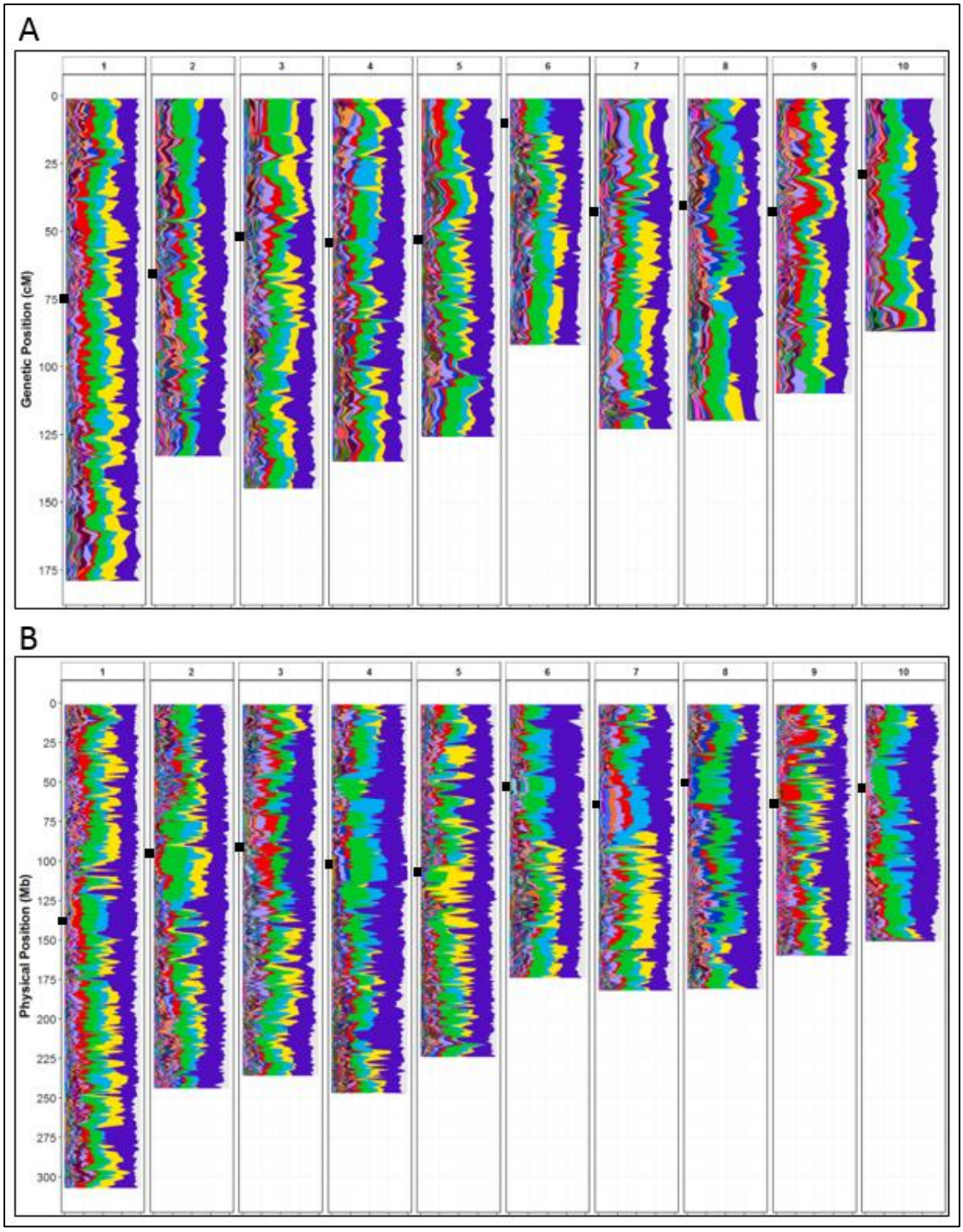

Figure A 2.3. Bacon plots showing composite IBS haplotypes across Monsanto ex-PVPs in genetic and physical space. Black squares mark the approximate centromere positions on each chromosome.

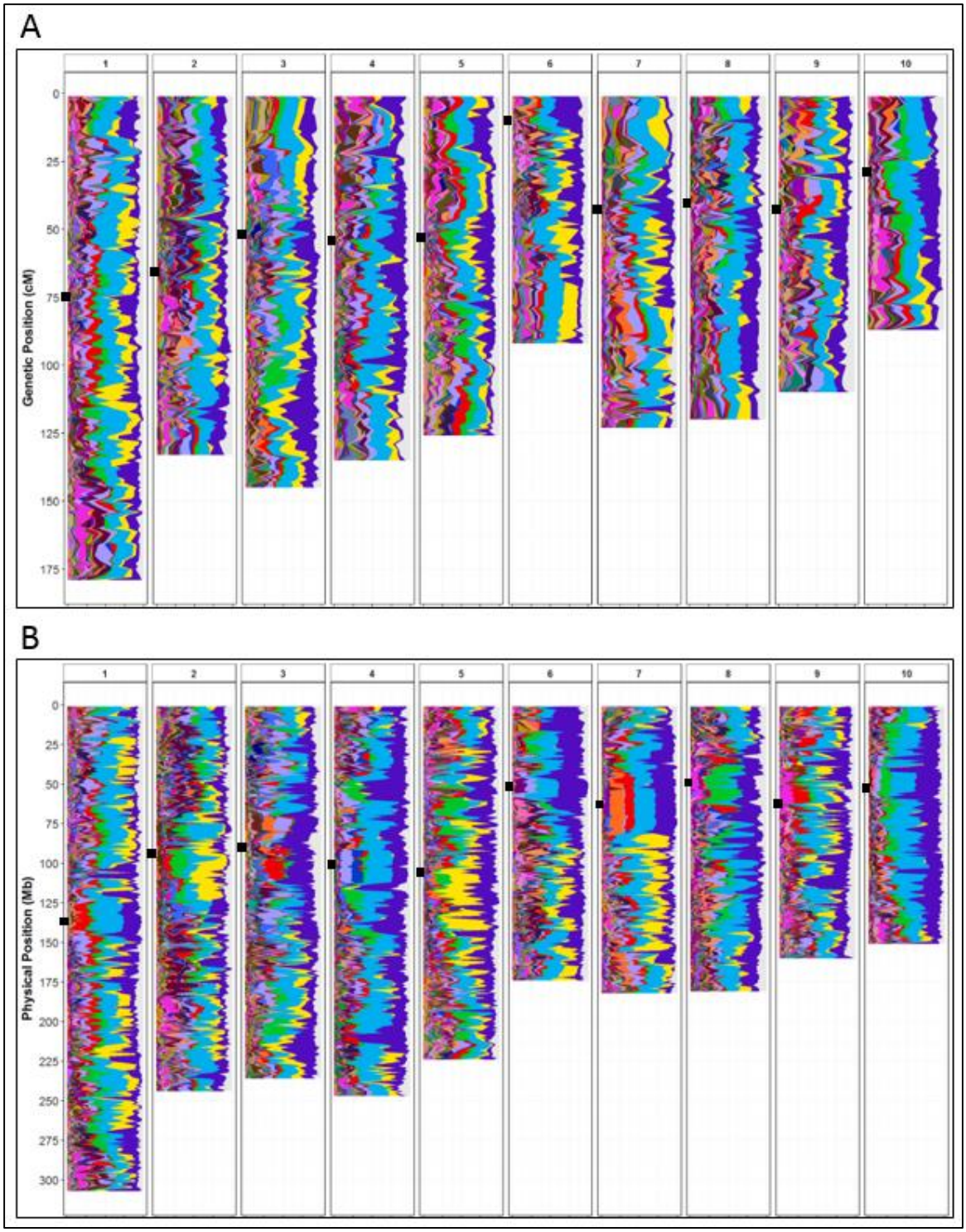

Figure A 2.4. Bacon plots showing composite IBS haplotypes across DowDuPont ex-PVPs in genetic and physical space. Black squares mark the approximate centromere positions on each chromosome.

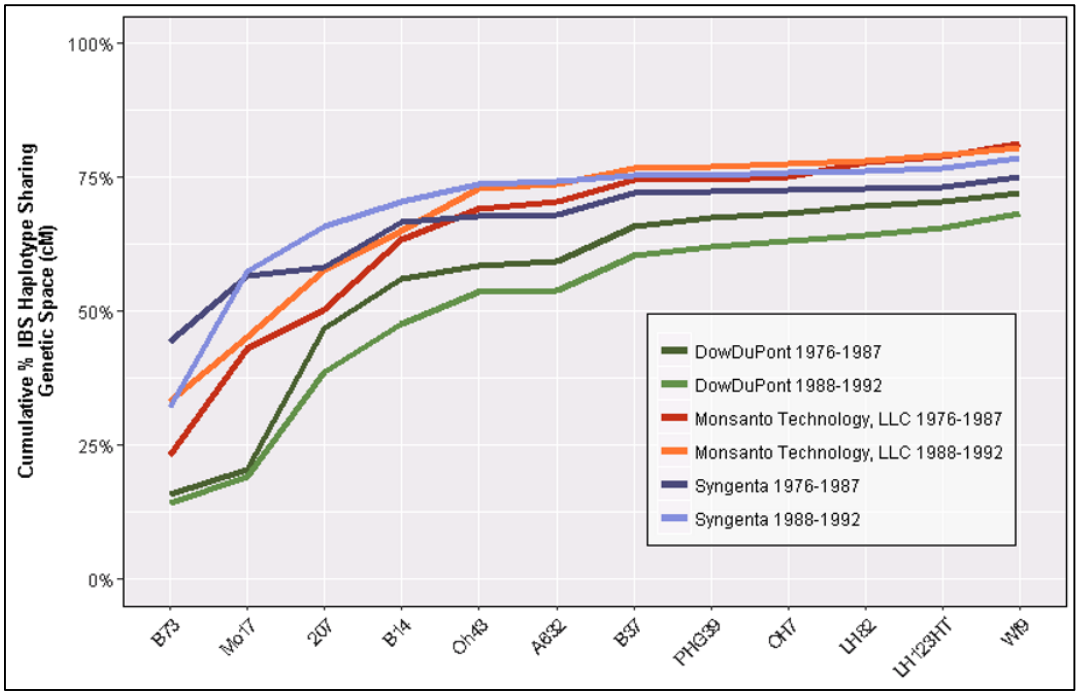

Figure A 2.5. Cumulative IBS haplotype sharing between ex-PVPs for DowDuPont, Monsanto and Syngenta. In addition to company, ex-PVPs are broken out into 2 groups based on application year: 1976-1987 and 1988-1992.

CHAPTER 3. A TOOL FOR VISUALIZATION OF SNP-BASED HAPLOTYPES

Article in preparation for publication in *BMC Bioinformatics*

Stephanie M. Coffman[1,2], Carson M. Andorf[3], Matthew B. Hufford[4], Thomas Lübberstedt[5]

Abstract

Background

Visualizing haplotype structure is a useful way for researchers to assess population structure diversity and relatedness among samples. While several tools exist to visualize haplotypes, they are primarily focused on small genomic regions rather than entire chromosomes.

Results

To demonstrate broad-scale haplotype visualization, we developed an interactive tool using the R Shiny framework. Three visualization panels allow a user to visualize haplotypes across the genome of a single individual, multiple individuals or compare haplotypes for multiple individuals against a reference sample.

Conclusions

The ShinyApp presented here demonstrates an easy-to-use, customizable interface for genome-wide haplotype visualization enabling researchers to quickly review haplotype structure and diversity across samples. The ShinyApp is freely available at https://github.com/scoffman/shiny_haplotype_vis.

[1] Primary Author
[2] Systems and Innovation for Breeding and Seed Products, DowDuPont, Johnston, Iowa
[3] USDA-ARS Corn Insects and Crop Genetics Research Unit, Iowa State University, Ames, Iowa
[4] Department of Ecology, Evolution & Organismal Biology, Iowa State University, Ames, Iowa
[5] Department of Agronomy, Iowa State University, Ames, Iowa

Keywords

haplotype, haplotype structure, interactive visualization, shiny

Background

A haplotype can be defined as a set of linked alleles which are inherited together. Haplotypes can provide additional information compared to single-nucleotide polymorphisms (SNPs) because there are more possible combinations of alleles. Haplotypes across individuals are identical-by-state (IBS) when they contain the same allelic information. Haplotypes can improve accuracy of genome-wide association studies (GWAS), genomic prediction (Ferdosi et al., 2016; Schrag et al., 2007) and mapping of quantitative trait loci (QTL; Kebede et al., 2016; Lu et al., 2010). Use of haplotype information can provide a more accurate assessment of population structure and depiction of relationships between individual samples (Fang et al., 2014; Gattepaille & Jakobsson, 2012; Haasl & Payseur, 2011; Lawson et al., 2012; Ralph & Coop, 2013).

Haplotypes are typically generated using the underlying SNP data of the genomes of a collection of individuals through a linkage disequilibrium (LD) or window-based approach. In a LD-based approach, an algorithm seeks to find the natural recombination breakpoints among haplotypes in a population. A window-based approach divides the genome into equal sized windows and haplotype groups are identified within each window. Haplotype groups can be generated from simple clustering and similarity algorithms (Gusev et al., 2009; Purcell et al., 2007; Swarts et al., 2014; Ward, 1963). Sets of ordered SNP alleles that cluster based on a similarity threshold in a genomic region are considered identical-by-state (IBS). Alternatively, more complex probability-based approaches which incorporate recombination data can also be used (Browning & Browning, 2009; Daly et al., 2001).

Population structure is frequently assessed through principal components analysis (PCA) and generation of STRUCTURE and fastStructure plots (Pritchard et al., 2000; Raj et al., 2014). Color coding of SNP alleles with tools such as Flapjack (Milne et al., 2010) can provide insight into the haplotype structure of a group of individuals. Haploview (Barrett et al., 2005) and GEVALT (Davidovich et al., 2007) enable LD-based haplotype construction and identification of tag SNPs useful in downstream analyses but do not provide context to the haplotypes in terms of visualization. The inPHAP (Jäger et al., 2014) interface enables interactive aggregation of single-nucleotide variant (SNV) data and visualization of region-based haplotypes using user-supplied metadata. Haplostrips (Marnetto et al., 2017) is a command-line tool that clusters SNP data into haplotypes based on a region of interest and generates simple plots of the resulting haplotype structure. Wang et al. (2012) developed a web-based genome browser to facilitate exploration of a mouse dataset and provides numerous tracks including haplotype count and haplotype diversity measures for exploring both smaller and larger genomic regions. HaploForge (Tekman et al., 2017) is modeled after HaploPainter (Thiele & Nürnberg, 2005) and generates pedigree-based haplotype visualizations for user-defined regions. These tools described often rely on SNP data as the input and with high density SNP data on many individuals, file sizes can become cumbersome. Further, users are typically limited to exploration of specific genomic segments at a time rather than a whole genome. Tools which emphasize visualization of haplotypes along with the underlying SNP data in specific regions of the genome are beneficial in analyses such as candidate gene identification but are not ideal for identifying broad-scale haplotype patterns across the genome.

Web-based, interactive graphics generated through libraries and platforms such as D3 (https://d3js.org/) and Shiny (https://shiny.rstudio.com/) are becoming more commonplace and have been used in analysis of genetic data (Pavlopoulos et al., 2015). Here, our objectives were to i) present an interactive interface created with the R Shiny framework for species-agnostic visualization of genome-wide haplotype structure and broad-scale comparison of IBS haplotypes across individuals and ii) demonstrate proof of concept for the tool using a small haplotype dataset for a maize inbred and three generations of ancestors.

Methods

The tool interface was developed using a web application framework for R called Shiny (version 1.0.5). A ShinyApp can be ran locally through software such as RStudio (https://www.rstudio.com/) or deployed on a server using Shiny Server. Our ShinyApp interface was built from three files: ui.R, server.R and styles.css. Overall styling of the app is defined in the styles.css. Layout of each panel is defined in the ui.R. Plot formatting and data manipulation based on the user input selections are controlled within server.R. Plots are generated using the 'ggplot2' package (Wickham, 2009). Input data are handled and displayed based on the functions and settings implemented by these scripts.

The graphical user interface (GUI)

The GUI can be launched locally by opening either the ui.R or the server.R file in RStudio and clicking 'Run App'. This interface consists of three panels: 'Single Samples View', 'Multiple Samples View' and 'Compare Samples to Reference'. Each panel contains a sample selection menu and options for viewing chromosomes using physical or genetic coordinates. The first panel allows for a genome-wide view of a single sample at a time. The other two panels allow selection of multiple samples and in the case of the third panel comparison of multiple samples to a single reference sample. The first and second panel

display the haplotype groups assigned to various genomic regions in a sample or samples. For this study, a sample is defined as one replicate of haplotype data for an individual maize line. Different colors are used to indicate different haplotype groups. The legend on these two panels contains the most frequent haplotype groups based on physical genome space.

Haplotype data source

The haplotype data were generated by running the FILLINFindHaplotypesPlugin (Swarts et al., 2014) in Tassel 5.0 (Bradbury et al., 2007) on a set of 11.3 million SNPs and 212 samples from the Maize HapMap3.2.1 (Bukowski et al., 2018). The samples consisted of 157 maize lines with expired Plant Variety Protection (ex-PVP; USDA, 2013) and 55 public maize lines. The 55 public maize lines included key founders which were important in the maize seed industry companies which generated the ex-PVP lines. The FILLINFindHaplotypesPlugin uses a non-overlapping SNP window-based approach. Here we used a window size of 2,000 SNPs. Each SNP window will be referred to as a haplotype block. Within each haplotype block, samples were grouped together based on similarity. This resulted in one or more haplotype groups for each haplotype block. For this study, a maximum diversity of 3% among samples within a haplotype group was allowed.

For demonstration purposes, a subset of 7 samples from the 212 will be summarized in the haplotype visualizations. These samples are the maize inbred lines LH205, LH74, LH119, A632, B73, Mt42 and B14. Together these inbreds form the pedigree structure for LH205 going back 3 generations (Figure 3.1). LH205 is an ex-PVP inbred line (PVP No. 9000049) which was developed by Monsanto.

Data formats and structures

The FILLINFindHaplotypesPlugin generates output files that contain the haplotype group information for each SNP window or haplotype block. These files were combined into

a simpler, single file format in R that could be utilized in the Shiny app. These combined haplotype data were exported as a 'feather' file using the feather package in R (Wickham, 2016). The feather file format is a binary format, optimized for high read and write performance. It is compatible with multiple scripting languages including R and Python. Use of feather format enables fast loading of the haplotype data when the Shiny app is launched. This also means that large haplotype datasets may be used with minimal impact to the launch time of the app. The general format for the combined haplotype data input is shown in Table 3.1. The haplotype data contain information on which samples are assigned to which haplotype group within each haplotype block. Not every sample must be assigned to a haplotype group in every haplotype block.

In addition to the haplotype data, a map positions file and sample metadata file are incorporated. The map positions file allows for bringing in genetic coordinates relative to physical coordinates. Sample metadata such as pedigree information can be provided in the sample metadata file for display in the 'Single Samples View'. The format for the map positions file can be seen in Table 3.2 and the sample metadata in Table 3.3.

Hierarchy for color coding haplotypes

The haplotype output from the FILLINFindHaplotypes plugin contains information on the haplotype groups in each haplotype block. Each haplotype group is assigned a name by the plugin based on the first sample that was placed into the group. To improve the ability to visually interpret the haplotypes, the haplotype groups were renamed based on a hierarchy. The hierarchy logic is described in detail in Figure 3.2.

The hierarchy used was simply an ordered list of samples with the key ancestors at the top. The hierarchy was applied to the haplotype data through an R script which renamed each haplotype group based on the sample in the haplotype group that was highest in the

hierarchy. For example, B73 was placed first in the hierarchy. Any haplotype group that contained B73 was re-named ‘B73’. The colors in the haplotype visualization directly reflect the haplotype group names.

Results

Haplotype Coloring Hierarchy

Recoding the haplotype group names based on a hierarchy improved visual interpretation of the haplotypes. In our small set of inbreds, we were interested in haplotype relationships between LH205 and its pedigree-based ancestors. The haplotype coloring that resulted from applying the color hierarchy matched closely with expected inheritance patterns of two-way cross inbreds. Large uninterrupted haplotype blocks were able to be related to direct ancestors.

Visualization of haplotype assignments for a single sample

Whole genome haplotype structure for the seven maize inbreds was visualized in the ‘Single Sample View’. The haplotype coloring observed for LH205 (Figure 3.3) showed heavy presence of B73 and B14 haplotypes. B73 and B14 belong to a breeding group termed ‘stiff-stalk’ (Eberhart et al., 1973). Maize inbreds in this breeding group are typically used as females in hybrid crosses. Presence of B73 and B14 haplotypes in LH205 is in line with the pedigree information and could indicate that LH205 is also a stiff-stalk line. The entirety of chromosome 2 for LH205 except for two small segments near 60cM and 70cM was IBS with B73. These small instances of differing haplotypes within larger solid colored blocks such as this could result from shared haplotypes between inbreds higher in the coloring hierarchy than LH205. Differing SNP densities available in each window when the haplotype groups were generated could also be a factor. Another region on chromosome 1 from 50-160cM showed LH205 is predominantly IBS with B14 but also contained some small segments that

were IBS with B73. These small segments are likely regions where B14 and B73 are IBS with one another. Because B73 is higher in the haplotype coloring hierarchy than B14, these segments were colored to match B73. In addition to B73 and B14, LH205 shared IBS haplotypes with other key maize founders such as Mo17 and Oh43.

Visualization of haplotype assignments across multiple samples

In the 'Multiple Samples View', haplotype structure for the inbreds in the LH205 pedigree were visualized side-by-side (Figure 3.4). Because the informed hierarchy-based haplotype coloring was performed on all samples, inferences could be made easily across samples in the visualization. Multiple samples in the LH205 pedigree structure showed shared haplotype structure. LH74 and LH119 share B73 as a parent and both of these lines share haplotypes with B73 at the top of chromosome 1. LH205, the progeny line of LH74 and LH119, also shares this haplotype. The ancestor Mt42 showed a very different haplotype structure compared to the other lines. Mt42 primarily contained haplotypes unique to Mt42 interspersed with several segments that are IBS with inbreds higher in the hierarchy than Mt42.

Comparison of haplotype assignments against a reference sample

The haplotypes for LH205 and its ancestors were recolored in the 'Compare Samples to Reference' tab to visualize a direct comparison of haplotypes in the ancestors to those present in LH205 (Figure 3.5). This view simplified comparison of shared haplotypes across samples by using a reference sample. Because the pedigree structure of LH205 is known, it was easy to see which segments were likely inherited from its parents LH74 and LH119. These segments could be further traced back to the other ancestors. In the region on chromosome 1 from ~50-160cM, LH205 matched to its parent LH74. In this same region, LH74 matched to its parent A632 and in turn, A632 matched to its parent B14. Tracing this

segment directly back to B14 in this view confirmed the large B14 segment observed on chromosome 1 for LH205 in the other views.

Discussion

Here, we demonstrated how the Shiny interface can be set up to plot haplotype structure across a group of individuals and provide impactful, interactive visualizations which are useful in characterization of germplasm. These visualizations capture a broader view of the population and help the researcher understand overall inheritance and relatedness. Shared genomic regions across a group of individuals, haplotype diversity and inferred inheritance through tracing haplotypes up a known pedigree can be quickly assessed.

In our example of the maize inbred LH205 and its ancestors, we were able to assess genome-wide haplotype structure of individual lines and make comparisons across lines. We identified ‘stiff-stalk’ as the likely breeding group for LH205, based on its shared haplotypes with the stiff-stalk lines B73 and B14. We identified shared regions across lines and regions which were unique to individual lines such as those present in Mt42. By comparing the haplotypes of individual ancestor lines directly to LH205, we were able to identify genomic segments that could be traced up the pedigree through the haplotype data. In the chromosome 1 region from ~50-160cM, we were able to identify that LH205 is not only IBS with B14 in this region but is also identical-by-descent (IBD) with B14.

Using a hierarchy-based approach to color the haplotypes improved the intuitiveness of our haplotype results. Recoding the haplotype group names based on a hierarchy was particularly beneficial because of the degree of relatedness among the individuals in the dataset. Because we were interested in pedigree-based ancestral composition in our set of samples, our coloring hierarchy placed the key ancestors at the top of the hierarchy, thereby resulting in more solid contiguous blocks that could be easily related to the ancestors.

Modifying haplotype group names using a color hierarchy is not required for the tool to run and using a color hierarchy does not result in a loss of information, however, a color hierarchy based on known information can improve ease of interpretation in the visualizations.

The availability of web-based tools which can be used and easily customized facilitates the process of exploring population structure. Publicly available haplotype visualization tools are typically limited to visualization of specific regions and often require input of SNP data. This can be ideal if the researcher is interested in haplotypes and variants in candidate genes or small genomic regions, but it makes assessment of genome level population structure challenging. Our tool demonstrates interactive visualization of haplotype data at the whole-genome level which is a feature not addressed by current tools. This enables whole-genome assessment of haplotype structure and visual identification of common and unique haplotype patterns across a set of individuals. Tools which perform haplotype visualization often require SNP calls as the input (Jäger et al., 2014; Marnetto et al., 2017). Because our tool requires haplotype information rather than the SNP calls, input file sizes are reduced. Further, use of haplotype groups as input rather than SNP data enables the researcher to load haplotype data generated from a haplotyping method of their choice.

Future work

The tool, as it stands, is effective for quickly visualizing haplotype structure in individual samples and making comparisons across samples. Future improvements could include loading custom sample lists and enabling custom sorts in the plot display. Region-specific visualization could be implemented through interactive zoom or user-entered start and stop positions. Additional options to customize the plots such as width, height and axis labeling could be added. Currently, there are 2 metadata columns that are displayed in the

plot titles shown in the ‘Single Sample View’. These metadata could be expanded and utilized for plotting of grouped samples. Use of a container program such as Docker has not been explored but would remove complications that might be experienced due to differing R package versions. Percentages of genome space accounted for by each haplotype group in the sample(s) displayed could be computed and displayed based on the input haplotype data. The tool does not provide an interactive haplotype similarity matrix or clustering of samples which may assist in interpreting the haplotypes observed in the population.

Work are ongoing to provide haplotype information through a Practical Haplotype Graph framework in multiple crops including maize (Johnson et al., 2018). The output would provide haplotype data and imputed SNP calls on samples that receive low coverage sequencing. A collaboration with MaizeGDB (https://www.maizegdb.org/) is underway to implement a web-based version of the haplotype visualization tool and to explore its potential extension to output from the Practical Haplotype Graph.

Conclusions

Here, we show an example of an interactive, web-based tool that allows broad-scale assessment of population structure given haplotype data. Although we used a set of maize inbreds as an example, the tool and the concept can easily be extended to other haplotype datasets and species particularly where long-range haplotype sharing is expected. These visualizations make it possible to quickly identify regions of commonality or difference and relate regions to a pre-defined list of ancestors. Identification of such regions can be useful in germplasm exploration and connecting haplotypes to traits.

List of abbreviations

Declarations

Ethics and approval and consent to participate

Not applicable.

Consent for publication

Not applicable.

Availability of data and material

The datasets used in this study are available at https://github.com/scoffman/shiny_haplotype_vis.

Competing interests

The authors declare that they have no competing interests.

Funding

The author SMC is employed by DowDuPont. The funder provided support in the form of salary and graduate program support for author SMC, but did not have any additional role in the study design, data collection and analysis, decision to publish or preparation of the manuscript.

Authors' contributions

SMC designed and ran the analyses, interpreted the results, developed the visualization tool and wrote the manuscript. MBH, CMA and TH assisted with the study design, interpretation of results and provided critical feedback during manuscript preparation.

Acknowledgements

The author SMC wishes to thank Justin Gerke and Dean Podlich for helpful discussions in preparation of this manuscript.

Endnotes

Figures

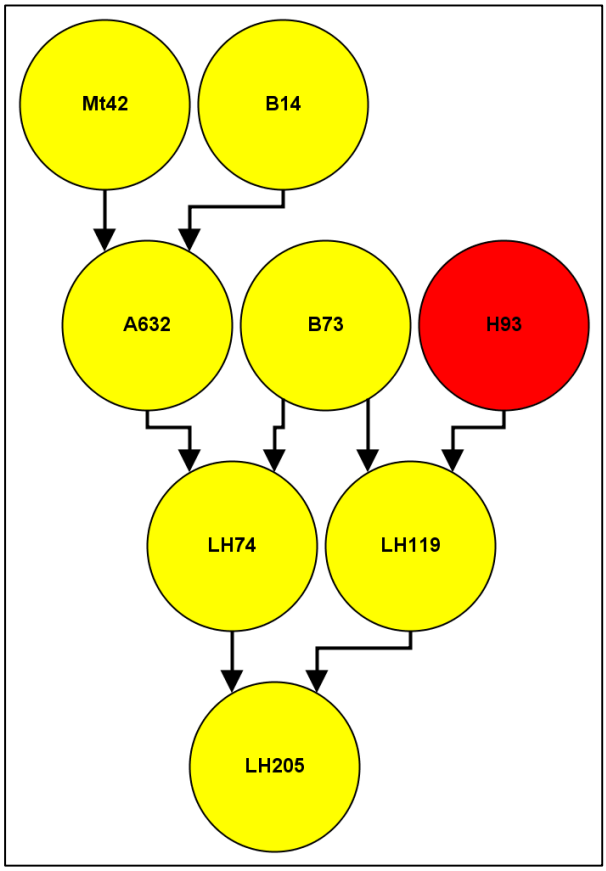

Figure 3.1. LH205 pedigree visualization. This pedigree was visualized with Helium (Shaw et al., 2014). LH205 is the most recent inbred and is placed at the bottom of the pedigree. Each level above LH205 is one successively one generation back in its pedigree. Yellow circles denote the inbred has haplotype information. Red circles denote that the inbred does not have haplotype information.

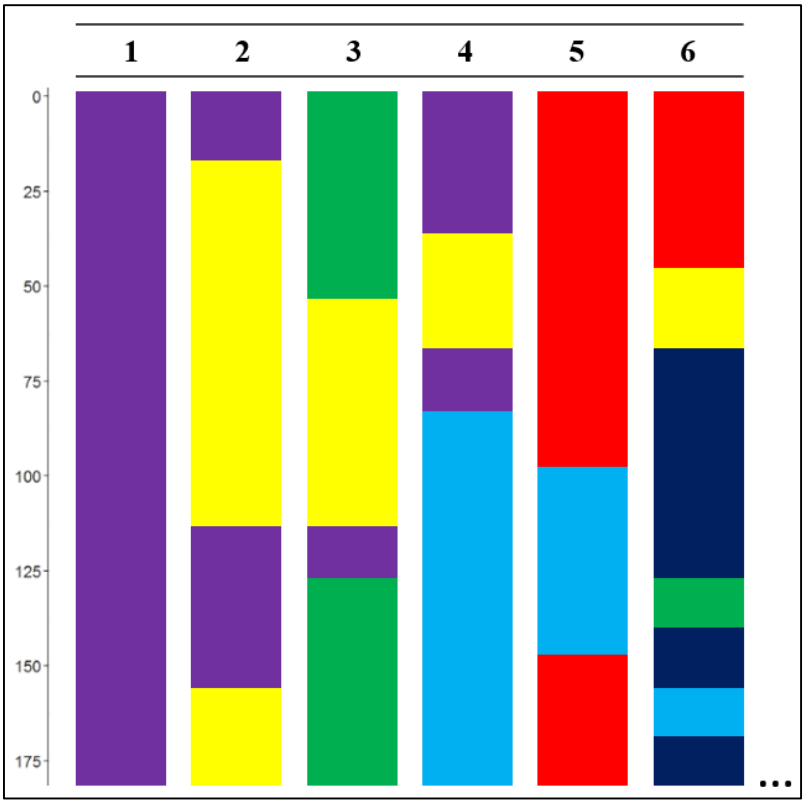

Figure 3.2. Hierarchy-based haplotype coloring logic. Here, a single chromosome is shown for six samples numbered 1 to 6 across the top. The y-axis is a genetic position scale in cM. The genomic segments are colored according to a hierarchy determined by the user. The hierarchy is simply an ordered list of the sample names. All of the haplotype group assignments for the first sample in the hierarchy get re-named to the sample name (i.e. named after itself). This is directly reflected in the coloring. Sample #1 will display one color across the entire genome (purple). Any genomic segments in any sample below sample #1 in the hierarchy that share the same haplotype group as sample #1 in a given haplotype block will also show purple. For example, haplotype blocks where sample #2 was not placed in the same haplotype group as sample #1 receive a new color (yellow). Similarly, any haplotype blocks where sample 3 shares haplotype groups with sample #1 or sample #2 will be colored accordingly. Any haplotype blocks where sample #3 does not share a haplotype group with sample #1 nor sample #2 will get a new color (green). This logic continues from left to right.

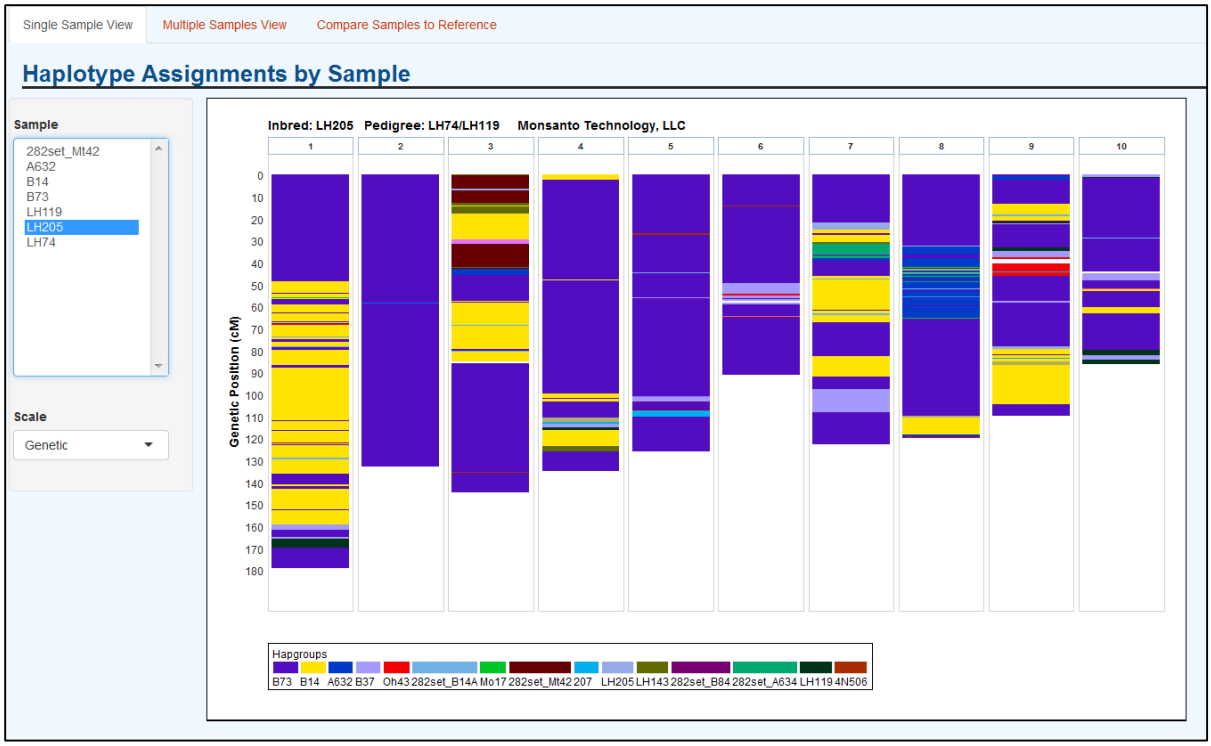

Figure 3.3. IBS haplotype assignments for LH205 across all 10 chromosomes in genetic space. Each column in the plot is a chromosome placed in ascending order from left to right. The y-axis displays genetic positions (cM). The two metadata columns provided for this sample are displayed next to the inbred name in the plot title. The most common haplotype groups observed in this sample, based on physical space, for this inbred are displayed in the legend at the bottom. Any SNP windows that did not receive a haplotype group assignment are colored white.

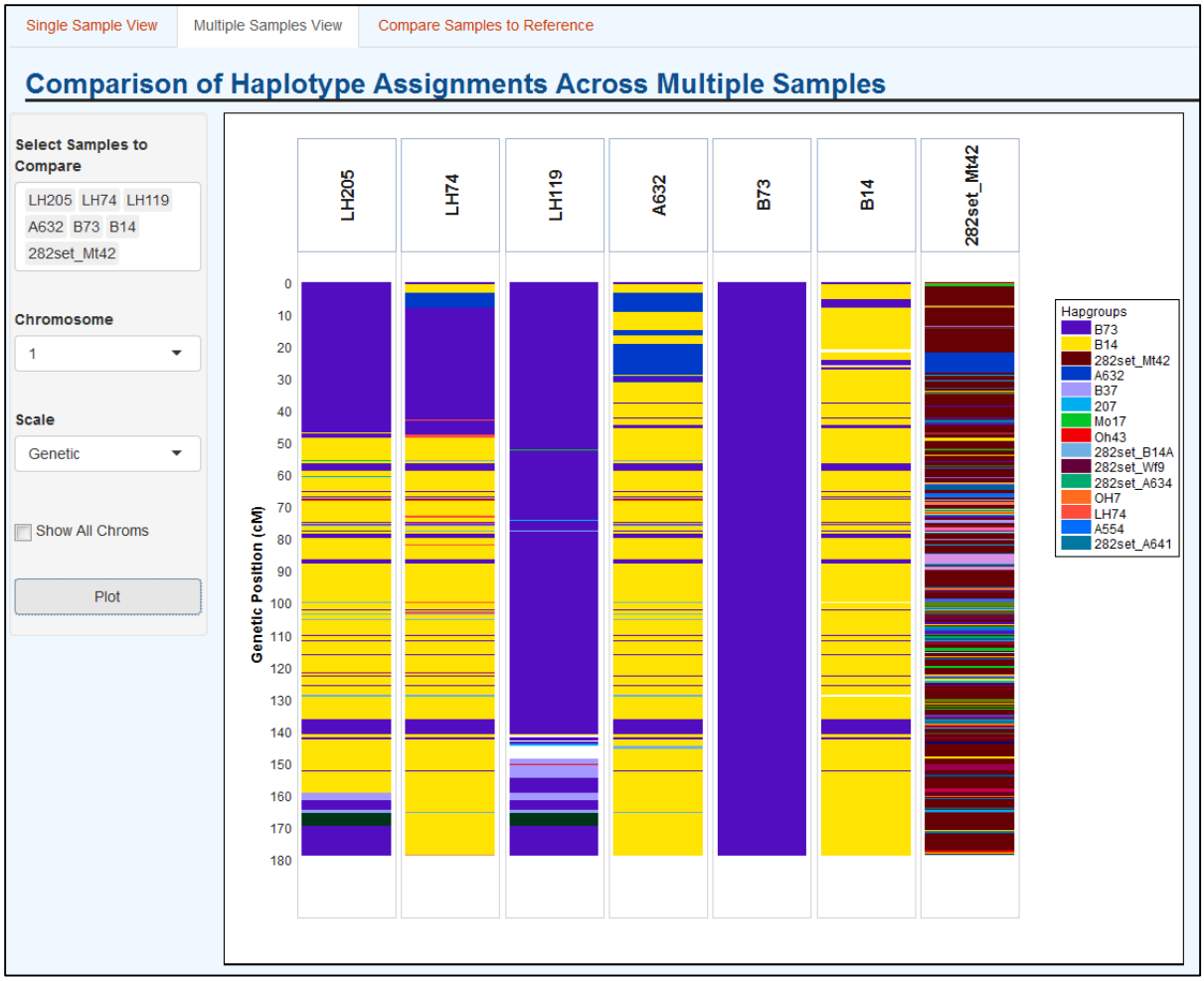

Figure 3.4. Comparison of the haplotype assignments on chromosome 1 in genetic space for LH205 and 6 of its ancestors. Each column represents a single chromosome for a single sample. The samples are ordered from left to right in the order in which they were entered in the selection menu. The most common haplotype groups across the selected samples and chromosome are displayed in the legend on the right. Just as in the 'Single Samples View', SNP windows which did not receive a haplotype group assignment are colored white.

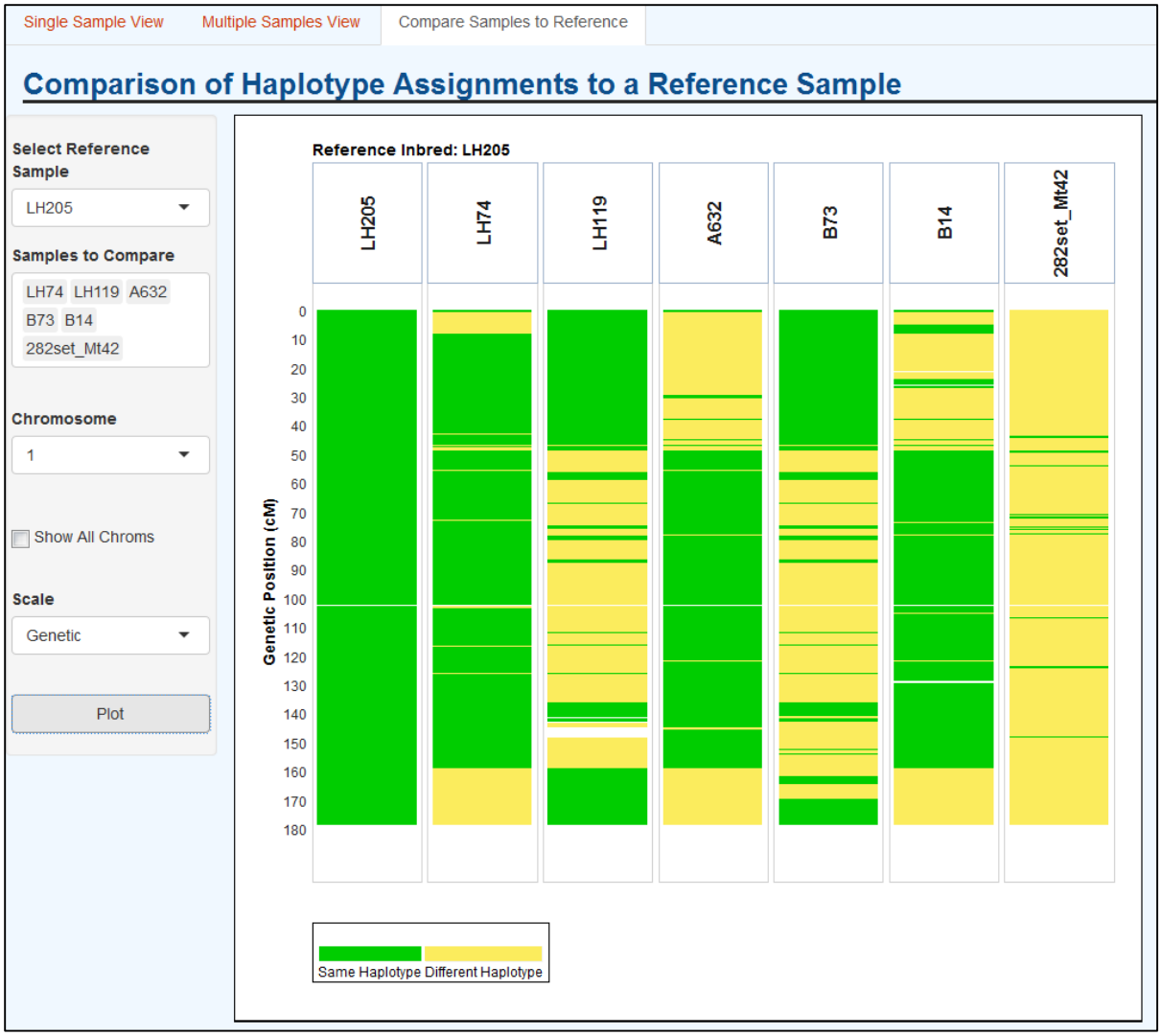

Figure 3.5. Comparison of haplotype assignments using LH205 as a reference inbred. Each column is a single chromosome for a single sample. Samples are ordered left to right with the reference inbred first, followed by the samples in the order in which they were entered in the sample selection menu. The reference inbred, LH205, is colored green. Instances where a sample is in the same haplotype group as LH205 in a given haplotype block are colored green. If the sample is in a different haplotype group, then it appears yellow for that haplotype block. As in the other views, haplotype blocks which did not receive a haplotype group assignment for a sample are colored white.

Tables

Table 3.1. Input haplotype file format. The first 20 rows of the input haplotype file used for the visualizations summarized are shown. All headers are required except for ‘block’. Each haplotype block has a corresponding ‘chr’, ‘startpos’ and ‘endpos’. Within each haplotype block there are haplotype groups. These are labeled as ‘origin_sample’ and each ‘origin_sample’ gets a unique color for display in the visualizations. To each haplotype group belong the samples in the ‘sample’ column. For example, the first haplotype block on chromosome 1 has three haplotype groups: ‘282set_Mt42’, ‘B14’ and ‘B73’. The ‘B73’ and ‘B14’ haplotype groups have three samples each (see ‘sample’ column). This haplotype file is saved as a ‘feather’ file which is easy to convert to from a comma-separated or tab-delimited file format through the ‘feather’ library in R or Python.

chr	block	sample	startpos	endpos	origin_sample	Hex_Code
1	0	282set_Mt42	56073	652006	282set_Mt42	#660005
1	0	A632	56073	652006	B14	#FFE303
1	0	B14	56073	652006	B14	#FFE303
1	0	B73	56073	652006	B73	#520dc0
1	0	LH119	56073	652006	B73	#520dc0
1	0	LH205	56073	652006	B73	#520dc0
1	0	LH74	56073	652006	B14	#FFE303
1	1	282set_Mt42	652074	921430	282set_Mt42	#660005
1	1	A632	652074	921430	B14	#FFE303
1	1	B14	652074	921430	B14	#FFE303
1	1	B73	652074	921430	B73	#520dc0
1	1	LH119	652074	921430	B73	#520dc0
1	1	LH205	652074	921430	B73	#520dc0
1	1	LH74	652074	921430	B14	#FFE303
1	2	282set_Mt42	921680	1289634	282set_Mt42	#660005
1	2	A632	921680	1289634	B14	#FFE303
1	2	B14	921680	1289634	B14	#FFE303
1	2	B73	921680	1289634	B73	#520dc0
1	2	LH119	921680	1289634	B73	#520dc0
1	2	LH205	921680	1289634	B73	#520dc0

Table 3.2. Map positions file format. The first 20 rows of the map positions file used in the visualization are shown. This file is required. All headers except ‘block’ are required. If genetic positions are not known, they can be filled in with 0’s. Genetic positions shown here are in cM.

chr	block	startpos	endpos	genpos_start	genpos_end
1	0	56073	652006	0.00	0.02
1	1	652074	921430	0.02	0.04
1	2	921680	1289634	0.04	0.05
1	3	1289635	1657834	0.05	0.06
1	4	1657911	1833287	0.06	0.13
1	5	1833309	2107624	0.13	0.55
1	6	2107626	2427644	0.55	0.72
1	7	2427645	2772733	0.72	0.83
1	8	2772737	2972908	0.83	1.24
1	9	2972935	3166085	1.24	2.24
1	10	3166099	3452948	2.24	3.54
1	11	3453000	3775964	3.54	5.33
1	12	3775966	3980731	5.33	6.38
1	13	3980752	4331327	6.38	7.26
1	14	4331329	4541531	7.26	8.15
1	15	4541536	4807178	8.15	9.38
1	16	4807179	5126364	9.38	10.36
1	17	5126366	5436348	10.36	12.33
1	18	5436350	5723761	12.33	13.77

Table 3.3. Input sample metadata file format. This file is required. All headers shown are required. Content only needs to be present in the 'sample' column. The 'pedigree' and 'comment' columns are only used in the plot title for the 'Single Sample View' tab.

sample	pedigree	comment
A632	(Mt42 x B14)$B14_3$	
B14		
B73	BSSS	
LH119	(H93/B73)/B73	Monsanto Technology, LLC
LH205	LH74/LH119	Monsanto Technology, LLC
LH74	A632/B73	Monsanto Technology, LLC
282set_Mt42		

References

Barrett, J. C., Fry, B., Maller, J., & Daly, M. J. (2005). Haploview: analysis and visualization of LD and haplotype maps. *Bioinformatics*, *21*(2), 263–265. https://doi.org/10.1093/bioinformatics/bth457

Bradbury, P. J., Zhang, Z., Kroon, D. E., Casstevens, T. M., Ramdoss, Y., & Buckler, E. S. (2007). TASSEL: software for association mapping of complex traits in diverse samples. *Bioinformatics*, *23*(19), 2633–2635. https://doi.org/10.1093/bioinformatics/btm308

Browning, B. L., & Browning, S. R. (2009). A Unified Approach to Genotype Imputation and Haplotype-Phase Inference for Large Data Sets of Trios and Unrelated Individuals. *The American Journal of Human Genetics*, *84*(2), 210–223. https://doi.org/10.1016/j.ajhg.2009.01.005

Bukowski, R., Guo, X., Lu, Y., Zou, C., He, B., Rong, Z., … Xu, Y. (2018). Construction of the third-generation Zea mays haplotype map. *GigaScience*, *7*(4). https://doi.org/10.1093/gigascience/gix134

Daly, M. J., Rioux, J. D., Schaffner, S. F., Hudson, T. J., & Lander, E. S. (2001). High-resolution haplotype structure in the human genome. *Nature Genetics*, *29*(2), 229–232. https://doi.org/10.1038/ng1001-229

Davidovich, O., Kimmel, G., & Shamir, R. (2007). GEVALT: An integrated software tool for genotype analysis. *BMC Bioinformatics*, *8*, 36. https://doi.org/10.1186/1471-2105-8-36

Eberhart, S. A., Debela, S., & Hallauer, A. R. (1973). Reciprocal Recurrent Selection in the BSSS and BSCB1 Maize Populations and Half-Sib Selection in BSSS 1. *Crop Science*, *13*(4), 451–456. https://doi.org/10.2135/cropsci1973.0011183X001300040017x

Fang, Z., Gonzales, A. M., Clegg, M. T., Smith, K. P., Muehlbauer, G. J., Steffenson, B. J., & Morrell, P. L. (2014). Two genomic regions contribute disproportionately to geographic differentiation in wild barley. *G3: Genes, Genomes, Genetics*, *4*(7), 1193–1203. https://doi.org/10.1534/g3.114.010561

Ferdosi, M. H., Henshall, J., & Tier, B. (2016). Study of the optimum haplotype length to build genomic relationship matrices. *Genetics Selection Evolution*, *48*, 75. https://doi.org/10.1186/s12711-016-0253-6

Gattepaille, L. M., & Jakobsson, M. (2012). Combining Markers into Haplotypes Can Improve Population Structure Inference. *Genetics*, *190*(1), 159–174. https://doi.org/10.1534/genetics.111.131136

Gusev, A., Lowe, J. K., Stoffel, M., Daly, M. J., Altshuler, D., Breslow, J. L., … Pe'er, I. (2009). Whole population, genome-wide mapping of hidden relatedness. *Genome Research*, *19*(2), 318–326. https://doi.org/10.1101/gr.081398.108

Haasl, R. J., & Payseur, B. A. (2011). Multi-locus inference of population structure: a comparison between single nucleotide polymorphisms and microsatellites. *Heredity*, *106*(1), 158–171. https://doi.org/10.1038/hdy.2010.21

Jäger, G., Peltzer, A., & Nieselt, K. (2014). inPHAP: Interactive visualization of genotype and phased haplotype data. *BMC Bioinformatics*, *15*, 200. https://doi.org/10.1186/1471-2105-15-200

Johnson, L. C., Bradbury, P., Casstevens, T., Ilut, D., Miller, Z., Punna, R., … Buckler, E. (2018). *A Practical Haplotype Graph for Determining Genomic Sequence*. Poster presented at the Plant & Animal Genome Conference XXVI, San Diego, CA.

Kebede, A. Z., Woldemariam, T., Reid, L. M., & Harris, L. J. (2016). Quantitative trait loci mapping for Gibberella ear rot resistance and associated agronomic traits using genotyping-by-sequencing in maize. *Theoretical and Applied Genetics*, *129*(1), 17–29. https://doi.org/10.1007/s00122-015-2600-3

Lawson, D. J., Hellenthal, G., Myers, S., & Falush, D. (2012). Inference of Population Structure using Dense Haplotype Data. *PLOS Genetics*, *8*(1), e1002453. https://doi.org/10.1371/journal.pgen.1002453

Lu, Y., Zhang, S., Shah, T., Xie, C., Hao, Z., Li, X., … Xu, Y. (2010). Joint linkage–linkage disequilibrium mapping is a powerful approach to detecting quantitative trait loci underlying drought tolerance in maize. *Proceedings of the National Academy of Sciences*, *107*(45), 19585–19590. https://doi.org/10.1073/pnas.1006105107

Marnetto, D., Huerta-Sánchez, E., & Price, S. (2017). Haplostrips: revealing population structure through haplotype visualization. *Methods in Ecology and Evolution*, *8*(10), 1389–1392. https://doi.org/10.1111/2041-210X.12747

Milne, I., Shaw, P., Stephen, G., Bayer, M., Cardle, L., Thomas, W. T. B., … Marshall, D. (2010). Flapjack—graphical genotype visualization. *Bioinformatics*, *26*(24), 3133–3134. https://doi.org/10.1093/bioinformatics/btq580

Pavlopoulos, G. A., Malliarakis, D., Papanikolaou, N., Theodosiou, T., Enright, A. J., & Iliopoulos, I. (2015). Visualizing genome and systems biology: technologies, tools, implementation techniques and trends, past, present and future. *GigaScience*, *4*, 38. https://doi.org/10.1186/s13742-015-0077-2

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., … Sham, P. C. (2007). PLINK: a tool set for whole-genome association and population-based

linkage analyses. *American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Raj, A., Stephens, M., & Pritchard, J. K. (2014). fastSTRUCTURE: Variational Inference of Population Structure in Large SNP Data Sets. *Genetics*, *197*(2), 573–589. https://doi.org/10.1534/genetics.114.164350

Ralph, P., & Coop, G. (2013). The Geography of Recent Genetic Ancestry across Europe. *PLOS Biology*, *11*(5), e1001555. https://doi.org/10.1371/journal.pbio.1001555

Schrag, T. A., Maurer, H. P., Melchinger, A. E., Piepho, H.-P., Peleman, J., & Frisch, M. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theoretical and Applied Genetics*, *114*(8), 1345–1355. https://doi.org/10.1007/s00122-007-0521-5

Shaw, P. D., Graham, M., Kennedy, J., Milne, I., & Marshall, D. F. (2014). Helium: visualization of large scale plant pedigrees. *BMC Bioinformatics*, *15*, 259. https://doi.org/10.1186/1471-2105-15-259

Swarts, K., Li, H., Navarro, R., Alberto, J., An, D., Romay, M. C., … Bradbury, P. J. (2014). Novel Methods to Optimize Genotypic Imputation for Low-Coverage, Next-Generation Sequence Data in Crop Plants. *The Plant Genome*, *7*(3). https://doi.org/10.3835/plantgenome2014.05.0023

Tekman, M., Medlar, A., Mozere, M., Kleta, R., & Stanescu, H. (2017). HaploForge: a comprehensive pedigree drawing and haplotype visualization web application. *Bioinformatics*, *33*(24), 3871–3877. https://doi.org/10.1093/bioinformatics/btx510

Thiele, H., & Nürnberg, P. (2005). HaploPainter: a tool for drawing pedigrees with complex haplotypes. *Bioinformatics*, *21*(8), 1730–1732. https://doi.org/10.1093/bioinformatics/bth488

USDA. (2013, July). United States Department of Agriculture: Plant Variety Protection Act and Regulations and Rules of Practice. Agricultural Marketing Service, Washington, DC.

Wang, J. R., de Villena, F. P.-M., & McMillan, L. (2012). Comparative analysis and visualization of multiple collinear genomes. *BMC Bioinformatics*, *13*(3), S13. https://doi.org/10.1186/1471-2105-13-S3-S13

Ward, J. (1963). Hierarchical Grouping to Optimize an Objective Function. *Journal of the American Statistical Association*, *58*(301), 236–244. https://doi.org/10.1080/01621459.1963.10500845

Wickham, H. (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. Retrieved from http://ggplot2.org

Wickham, H. (2016). *feather: R Bindings to the Feather "API."* Retrieved from https://CRAN.R-project.org/package=feather

CHAPTER 4. GENERAL CONCLUSIONS

The general goal of this project was to employ haplotype sharing analysis to uncover haplotype structure and diversity among ex-PVPs and relate those haplotypes to key North American maize founder lines. The results presented build on previous studies by providing high resolution haplotype data which uncover specific regions of haplotype sharing between ex-PVPs and founders and the similarities and differences across seed industry companies and heterotic groups. In addition to gaining resolution in understanding the relationships between ex-PVPs and founders, we also discussed the importance of data integrity and summarized source variation that was observed during the sample filtering process. We also provided examples of haplotype sharing which occurs between specific founders and emphasized that these founders are not all completely unique. Haplotype sharing in a specific genomic segment between one ex-PVP and a specific founder does not necessarily mean that the ex-PVP cannot be related to any other founder line. Although not all current ex-PVPs were included for this study, we demonstrated how relating haplotypes present in ex-PVPs to key founders provides an intuitive way to understand the breeding history of industry germplasm. Further application of this haplotype data could be realized through study of specific haplotypes associated with traits, facilitating selection of ex-PVP lines for use in breeding programs and studies of diversity and heterosis and inference of haplotypes in PVPs not yet expired.

Interactive visualization of haplotype data was valuable in assessment of the results in chapter 2. Utilizing a subset of the overall haplotype data generated in chapter 2, we summarized an example of an interactive, web-based tool in chapter 3 that allows broad-scale assessment of population structure given haplotype data. The aim of developing the tool was

primarily to enable haplotype visualization for a set of maize inbreds. However, the tool is set up such that it can receive haplotype data from any haplotype dataset provided it is formatted correctly. This allows the tool to be extended not just to other haplotype datasets but other species. The tool enables unique views of haplotype sharing across individuals. These visualizations make it possible to quickly identify regions of commonality or difference and relate regions to a pre-defined list of ancestors or key lines the researcher wishes to make comparisons against. Identification of shared and differentiated regions can be useful in germplasm exploration and connecting haplotypes to traits. Work are ongoing with Maize GDB to utilize this tool as a model for visualization of other maize haplotype datasets.